Section F – Technology and Analytics

Introduction to Technology and Analytics

Section F constitutes 15% of the CMA Part 1 Exam. Section F includes *Information Systems*, *Data Governance*, *Technology-enabled Finance Transformation*, and *Data Analytics*.

- The focus of *Information Systems* is on accounting information systems, inputs to them, and the use of their outputs.
- *Data Governance* involves the management of an organization's data assets and data flows. The objective of data governance is to enable reliable and consistent data so that management is able to properly assess the organization's performance and make decisions.
- *Technology-enabled Finance Transformation* covers issues of importance for management accountants such as robotic process automation, artificial intelligence, cloud computing, and blockchains.
- *Data Analytics* also covers current technological and analytical issues that management accountants need to be familiar with. Business intelligence, data mining in large datasets, the use of statistics and regression analysis, and visualization of data by charting are covered.

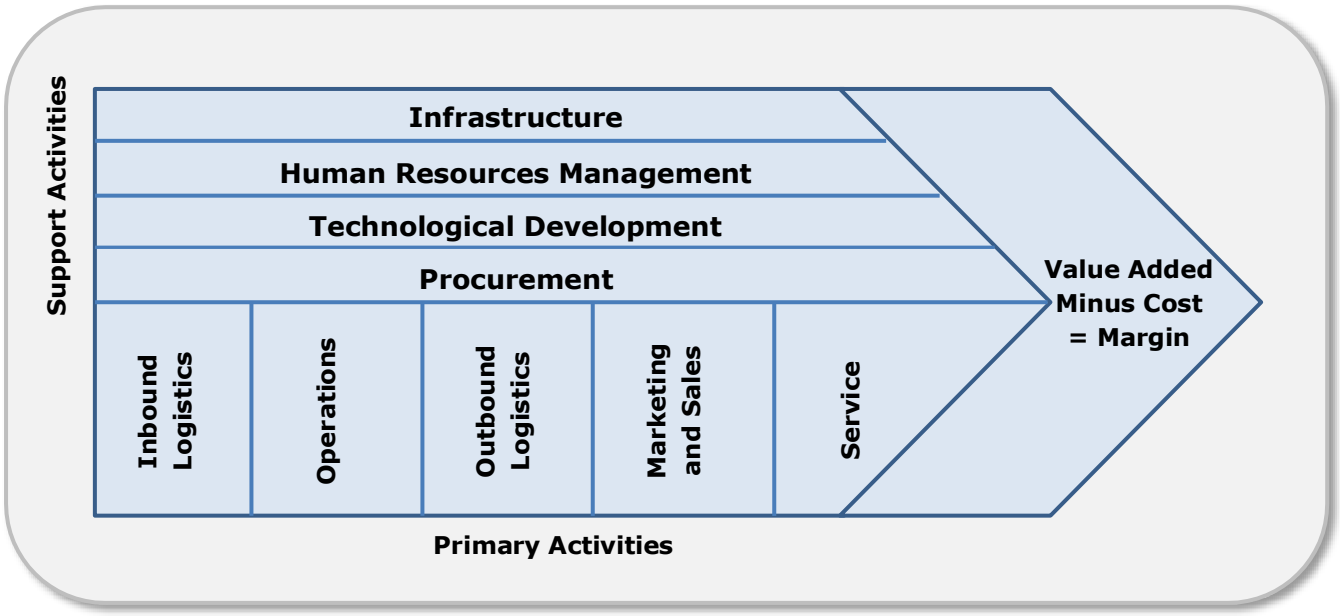
– Information Systems

The Value Chain and the Accounting Information System

The goal of any organization is to provide value to its customers. A firm's **value chain** is the set of business processes it uses to add value to its products and services.

The more value a company creates and provides to its customers, the more its customers will be willing to pay for its products or services, and the more likely they are to keep buying those products or services. A business will be profitable if the value it creates for its customers is greater than its cost of producing the products and services it offers. All of the activities in the value chain contain opportunities to increase the value to the customer or to decrease costs without decreasing the value the customer receives.

The value chain was discussed earlier in this volume. To review, the value chain as envisioned by Michael Porter looks like the following:



Though various organizations may view their value chains differently depending on their business models, the preceding graphic encompasses the business processes used by most organizations. The primary activities include the processes the organization performs in order to create, market, and deliver products and services to its customers and to support the customers with service after the sale. The support activities make it possible for the organization to perform the primary activities.

An organization's accounting information system (AIS) interacts with every process in the value chain. The AIS adds value to the organization by providing accurate and timely information so that all of the value chain activities can be performed efficiently and effectively. For example:

- Just-in-time manufacturing and raw materials inventory management is made possible by an accounting information system that provides up-to-date information about inventories of raw materials and their locations.
- Sales information can be used to optimize inventory levels at retail locations.
- An online retailer can use sales data to send emails to customers suggesting other items they might be interested in based on items they have already purchased.
- Allowing customers to access accounting information such as inventory levels and their own sales orders can reduce costs of interacting with customers and increase customer satisfaction.
- A variance report showing a large unfavorable variance in a cost indicates that investigation and possibly corrective action by management is needed.
- An AIS can provide other information that improves management decision-making. For instance:
    - It can store information about the results of previous decisions that can be used in making future decisions.
    - The information provided can assist management in choosing among alternative actions.

The Supply Chain and the Accounting Information System

Parts of a company's value chain are also parts of its **supply chain**. A company's supply chain describes the flow of goods, services, and information from the suppliers of materials and services to the organization all the way through to delivery of finished products to customers. In contrast to the value chain, the activities of a company's supply chain also take in outside organizations.

Nearly every product that reaches an end-user represents the coordinated efforts of several organizations. Suppliers provide components to manufacturers who in turn convert them into finished products that they ship to distributors for shipping to retailers for purchase by the consumer. All of the organizations involved in moving a product or service from suppliers to the end-user (the customer) are referred to collectively as the supply chain.

A well-designed accounting information system can improve the efficiency and effectiveness of a company's supply chain, thus enhancing the company's profitability.

Automated Accounting Information Systems (AIS)

Automated accounting information systems are computer-based systems that transform accounting data into information using the fundamental elements of paper-based accounting systems, but with electronic processing.

Elements of Automated Accounting Information Systems

Any financial accounting information system, automated or otherwise, captures transactions or business events that affect an entity's financial condition. The transactions are recorded in **journals** and **ledgers**. **Journals** are used to record accounting transactions. A journal contains a chronological record of the events

that have impacted the business's finances. Journal entries show all the information about a transaction, including the transaction date, the accounts debited and credited, and a brief description. Journal entries are then posted to the **general ledger**.

The general ledger contains a separate account for each type of transaction. The list of general ledger accounts used by an organization is called its **chart of accounts**. In a paper-based accounting system, only an account name is needed, but in an automated accounting information system, the accounts need to have numbers so that input is done in a consistent manner.

An automated accounting information system stores information in files. **Master files** store permanent information, such as general ledger account numbers and history or customer account numbers and historical data for each customer. **Transaction files** are used to update master files, and they store detailed information about business activities, such as detail about sales transactions or purchase of inventory. For each general ledger account number, the general ledger master file stores the transactions that have adjusted that account's balance along with some basic information such as date and source. The detail needed to maintain an audit trail is stored in transaction files. The detail in the transaction files may be needed in future years if it becomes necessary to adjust or restate prior period financial statements, so it is a good idea to maintain the transaction files for a number of years. After the transaction details are no longer needed, though, the transaction files can be deleted.

In an automated accounting information system, accounts in the general ledger chart of accounts are numbered using **block codes**. Block codes are sequential codes that have specific blocks of numbers reserved for specific uses. For example, a general ledger account numbering system is used to organize the accounts according to assets, liabilities, equity, incomes, and expenses. An entity can use any numbering scheme it wants, but one method of organizing account numbers uses account numbers beginning with "1" for assets, "2" for liabilities, "3" for equity, "4" for incomes, and "5" for expenses. Thus, if 4-digit account numbers are being used, then all asset accounts would be in the 1000 block, all liability accounts in the 2000 block, and so forth. The numbers following the number in the first position subdivide the types of accounts more finely. For example, in the asset section, current asset accounts might begin with "11" while noncurrent asset accounts begin with "12." Within the current asset section, then, cash might be 1110, accounts receivable might be 1120, and inventory might be 1130.

Special journals are used for specific kinds of transactions, and in a computerized system, the journals are known as **modules**. For example, an order entry module may be used to record sales. When a credit sale is recorded in the order entry module, the order entry module updates the customer's account in the accounts receivable module (to add the charge to the account); it updates the sales module (to increase the sales revenue); and it updates the inventory module (to reduce the items on hand by the items sold). It also updates the general ledger to record the increase to accounts receivable, the increase to sales revenue, the decrease in the value of inventory by the cost of the items sold, and the equivalent increase in cost of goods sold expense. If sales tax or value-added tax is invoiced, those items are recorded as well in the general ledger. When a customer pays, the receipt is recorded in the cash receipts module, which updates the customer's account in the accounts receivable module at the same time. The journal transactions update the general ledger to record the increase to the cash account and the decrease to the accounts receivable account.

In an automated accounting system, transactions are recorded electronically. They may be input by employees, but they may just as easily be recorded automatically. For example, a transaction may be created automatically in the order entry module when a customer places an order over the Internet.

When a transaction is created in a module, the input includes a **transaction code** that identifies, for example, a transaction in the order entry module as a sale transaction. The code causes the data entered with that transaction code to be recorded in the other modules and in the proper general ledger accounts. The specific transaction code for a sale may be set as the default code in the transaction code input field within the order entry module, although the code can be changed if instead, for example, a sales return is being processed.

Codes, both numeric and alphanumeric, are used elsewhere in an automated AIS, as well. For example, **sequence codes** are used to identify customer sales invoices created in the order entry module and may be used for new customer accounts in the accounts receivable master file. Sequence codes are simply assigned consecutively by the accounting system. For example, when a new customer account is added to the accounts receivable master file, the AIS may automatically assign the next consecutive unused account number to the new account. The number has no other meaning.

In a responsibility accounting system, a code is used to identify the responsibility center that is the source of the transaction, and that code is part of transactions input to the AIS, as well.

Example: General ledger expense account numbers used for advertising expenses include a code for the department that initiated the cost. The expense account number identifies the type of advertising medium, for example television advertising, followed by codes indicating the type of advertising expense and the product advertised.

Thus, a production expense for a television commercial advertising a specific kitchen appliance such as a blender is coded to the television advertising account for commercial production for blenders. As the cost is recorded in the AIS, the cost is directed to the correct responsibility center code and expense account, as follows:

120 = Advertising Department	5 = Expense	16 = Advertising	2 = Television	7 = Production Costs	31 = Blender
120	5	16	2	7	31

Thus, the full expense account number charged is 5162731 in department 120. That account number in that responsibility center accumulates only expenses for television advertising production costs for blender advertising that have been committed to by the advertising department.

As a result, the different types of advertising expenses are clearly delineated in the general ledger according to responsibility center, type of expense, advertising medium, type of cost, and product advertised, enabling easier analysis of the data.

Output of an Automated Accounting Information System

The data collected by an AIS is reported on internal reports. The internal reports are used by accountants to prepare adjusting entries, by management for analysis and decision-making, and by both accountants and management to produce the external financial reports needed.

An AIS needs to be designed so that it will be able to produce the reports that users will need. In the preceding example, for instance, before the expense codes were developed, management decided it needed to know not only total advertising expense, but also how much was spent for advertising in each advertising medium, within each medium how much was spent for advertising production and how much for media charges, and within those subdivisions, how much was spent to advertise each product.

Reports from an AIS may be paper reports, screen reports, or reports in various other forms such as audio reports. They could be regularly scheduled reports or they could be produced on demand. Good reports should have the following characteristics:

1) The report should include a date or dates. For example, a current customer list should show the date the report is "as of." A balance sheet should also show the "as of" date of the report. An income statement for a period of time such as the year to date should indicate the period covered by the report, such as "January 1 through April 30, 20X9."

2) The report should be consistent over time so managers can compare information from different time periods, consistent across segments so management can compare segment performance, and consistent with generally accepted accounting principles so the report can be understood and used.

3) The report should be in a convenient format and should contain useful information that is easy to identify. Summary reports should contain financial totals and comparative reports should provide related numbers such as actual versus budgeted amounts in adjacent columns.

Accounting Information System Cycles

Transaction cycles are grouped business processes for which the transactions are interrelated. They include:

- Revenue to cash cycle.
- Purchasing and expenditures cycle.
- Production cycle.
- Human resources and payroll cycle.
- Financing cycle.
- Fixed asset cycle (property, plant, and equipment).
- General ledger and reporting systems.

Revenue to Cash Cycle

The revenue to cash cycle involves activities related to the sale of goods and services and the collection of customers' cash payments. The cycle begins with a customer order and ends with the collection of the cash from the customer.

To include the collection of customers' cash payments for sales, the company needs to maintain accurate records of customers' outstanding invoices. Thus, the accounts receivable subsidiary ledger is an important function of the AIS. Customer records also include payment history, assigned credit limit, and credit rating.

The accounting information system is used for:

- Tracking sales of goods and services to customers.
- Recording the fulfilling of customer orders.
- Maintaining customer records.
- Billing for goods and services.
- Recording payments collected for goods and services provided.
- Forecasting sales and cash receipts using the outputs of the AIS.

The process of entering an order and sale into the AIS also updates the inventory module so that items sold are deducted from on-hand inventory and their costs charged to cost of goods sold.

The AIS should include a means to invoice customers as products are shipped. The production of the invoice and the production of the packing slip should occur simultaneously, and nothing should be shipped without a packing slip.

The AIS needs to be able to produce analytical reports of sales orders, sales terms, and payment histories for customers for use in predictive analytics. Sales orders can be used to predict future sales, and the sale terms can be used to make cash flow forecasts.

Activities in and inputs to the revenue to cash cycle include:

- Receipt of customer orders and creation of sales orders. The sales order serves as input to the AIS and contains full customer information, payment method, item or items sold, prices, and terms.
- Inventory is checked to determine whether the item or items ordered are in stock. A process needs to be in place to track and follow up on backordered items.
- An order confirmation may be emailed to the customer, particularly for orders received on the Internet.
- If the order includes a request for credit, the customer's credit is checked. A process needs to be in place to screen for fraudulent orders, as well, and much of that can be automated in the AIS.
- If the order is to be paid by credit card, preliminary credit card authorization is obtained for the amount due.
- Input is created to the AIS for the invoice and the packing slip. (If the order was received over the Internet, the input to the AIS may be automatically created.)
- Notice is sent to the warehouse to print the packing slip, pick the order, and prepare it for shipping. For a service, the service is scheduled.
- Shipping information is processed, the package is moved to the shipping dock to be picked up by the carrier, or the service is provided. For a credit card sale, the credit card charge is released. For a sale made on credit, the invoice is sent to the customer.
- The accounts receivable module, inventory module, and general ledger are updated and reports are processed.
- A process needs to be in place to handle approval and processing of returns. If a return is approved, the process needs to include making sure the returned item is physically received and, if it can be resold, that it is physically delivered back to the warehouse and added to inventory on hand in the inventory module at the correct cost. A credit is created to the customer's account and an appropriate journal entry is made to the general ledger.
- When payment is received for a credit sale, a remittance advice (usually detached from the invoice and sent back by the customer) should accompany the customer's payment. The remittance advice is used to input the information about the payment received to the correct customer account in the accounts receivable module, including the amount received and the invoice or invoices the payment is to be applied to. The AIS also updates the cash receipts journal and the general ledger to record the increase to cash and the decrease to accounts receivable.

Inputs to the revenue cycle can be made by desktop computer, but they may also be voice inputs, input with touch-tone telephones, or input by wireless capability using tablet computers. Sales made on the Internet may automatically update the subledgers and the general ledger in the AIS. Outside sales people may use laptops, mobile devices, portable bar code scanners, and other types of electronic input devices to enter sales orders.

Outputs of the revenue to cash cycle include:

- Internal management reports such as sales reports and inventory reports.
- Customer statements summarizing outstanding sales invoices by customer, payments received, and balance owed.
- Customer aging reports, showing the total accounts receivable outstanding balance broken down according to ranges of time outstanding, such as amounts outstanding for 0 to 30 days, 31 to 60 days, and so forth.
- Collection reports, showing specific accounts that need follow-up for overdue balances.

- Information from sales reports, receivables aging reports, and receipts reports is used as input to a cash receipts forecast.
- Various outputs are used in preparing closing entries and producing external financial statements. For example, the aging report and other information is used in estimating credit loss expense for the period.

Purchasing and Expenditures Cycle

The purchasing and expenditures process involves obtaining items and services in a timely manner at the lowest price consistent with the quality required, managing the inventory, and seeing that payment is made for items purchased and services received. The responsibilities of the purchasing function include:

- Locating and approving reputable vendors who offer quality goods and services at reasonable prices. Vendor shipping and billing policies, discount terms, and reliability are important concerns in approving vendors.
- Maintaining relationships with suppliers and other partners in the company's supply chain.

The purchasing and expenditures cycle begins with a request for goods or services and ends with payment to the vendor for the goods or to the provider of the service.

The accounting information system is used for:

- Tracking purchases of goods or services.
- Tracking amounts owed and making timely and accurate vendor payments.
- Maintaining vendor records and a list of authorized vendors.
- Managing inventory to ensure that all goods purchased are received, properly recorded in the AIS, and properly dispensed from inventory.
- Forecasting purchasing needs and cash outflows.

The inventory control function in the AIS interfaces with production departments, purchasing, vendors, and the receiving department.

Activities and inputs to the purchasing and expenditures cycle include:

- An internal purchase requisition is prepared by the requesting manager or department. Requests for raw materials may be automated in the AIS when inventories fall below pre-determined levels, or employees may key in the requests.
- The request is transmitted electronically to the purchasing department.
- The purchasing department selects the vendor, prepares input to the AIS for a purchase order, and transmits the purchase order to the vendor.
- The purchasing department also transmits the information from the purchase order to the company's receiving department so it will have the record of the order when the order is delivered. However, as a control, the information the receiving department has access to should not include the quantities ordered of each item. Thus, the receiving clerk must actually count and record the items received rather than simply assuming the quantities ordered were the quantities received.
- When items are received, the receiving department creates an electronic receiving report. The receiving report may be integrated in the AIS with the purchase order input as long as the receiving personnel cannot access the quantities ordered. In other words, the receiving clerk may create the receiving report by accessing each item on the purchase order electronically and inputting the quantity counted that was received, but the receiving clerk should not be able to see the quantity ordered.

- Recording the receipt of the items electronically in the AIS updates the inventory on hand in the inventory module, increases inventory in the general ledger by the amount of the cost, and creates a payable in the accounts payable module and in the general ledger.

  Note: Items shipped FOB shipping point belong to the purchaser as soon as they are shipped. Items shipped FOB shipping point that have been shipped but not yet received as of a financial statement date should be accrued as payables and the items should be included in ending inventory. The AIS should have a means of properly handling those transactions.

- The invoice received from the vendor is compared with the purchase order and the receiving report. A packing slip and a bill of lading from the freight carrier may also be received and they should be included in the review.
- A process should be in place to investigate any differences in the items, quantity, and prices between and among the purchase order, receiving report, invoice, packing slip, and bill of lading.
- The invoice information is input into the accounts payable module to complete the information that was added to the accounts payable module by the receiving report.
- The AIS should include controls that limit the potential for duplicate invoices to be paid. For example, if an invoice is input that matches one from the same vendor with the same invoice number or amount that has already been paid, it should be flagged for investigation before payment.
- If the item "received" was a service, approval should be received from a manager above the level of the requesting manager before payment is prepared and sent. The higher-level approval is a control to limit the opportunity for a manager to create a fictitious company, give fictitious service business to that company, and approve payment to be sent to that company (that is actually sent to him- or herself).
- If everything on the purchase order was received and if the items, quantities, and prices on the invoice match the items, quantities, and prices on the purchase order, payment is prepared in the AIS and sent. The payment may be sent as a paper check printed by the AIS. However, payments are increasingly being sent by electronic funds transfer (EFT) through the automated clearing house (ACH) whereby the funds are deducted from the purchaser's bank account and sent electronically to the vendor.
- The accounts payable module and the general ledger are updated and reports are processed.
- Petty cash or procurement cards may be used for smaller purchases.

Inputs to the purchasing and expenditures cycle can be made with desktop computers, bar code scanners, radio or video signals, magnetic ink characters on checks, scanned images, or entered into a tablet computer and transmitted wirelessly.

Outputs of the purchasing and expenditures cycle include:

- The check (if a paper check is used) or payment advice (for an ACH transfer) and the payment register.
- A cash requirements forecast.
- Discrepancy reports that note differences in items, quantities, or amounts on the purchase order, the receiving report, and the vendor's invoice or duplicate invoice numbers or duplicate amounts to a vendor.

As noted previously, the discrepancy report is needed to prevent authorization of payment to a vendor until any differences between or among the items, quantities, or prices on the purchase order, the receiving report, and the purchase invoice or any potential duplicate invoices have been investigated and resolved.

Using unpaid vendor invoices, outstanding purchase orders, and reports of items received for which the invoices have not yet been received, the AIS can predict future cash payments that will be needed and the dates they will be needed.

Production Cycle

The production cycle involves conversion of raw materials into finished goods and the goal is to do that as efficiently as possible. Computer-assisted design technology and robotics are often used.

The production process begins with a request for raw materials for the production process. It ends with the completion of manufacturing and the transfer of finished goods inventory to warehouses.

The accounting information system is used for:

- Tracking purchases of raw materials.
- Monitoring and controlling manufacturing costs.
- Managing and controlling inventories.
- Controlling and coordinating the production process.
- Providing input for budgets.
- Collecting cost accounting data for operational managers to use in making decisions.
- Providing information for manufacturing variance reports, usually using job costing, process costing, or activity-based costing systems.

Activities of and inputs to the production cycle include:

- Production managers issue materials requisition forms when they need to acquire raw material from the storeroom.
- Physical inventory is taken periodically and reconciled to inventory records. The number of items physically counted is input to the raw materials inventory module for reconciliation, and the accounting records are updated.
- If the level of raw materials inventory in the storeroom falls below a predetermined level, a purchase requisition is issued to the purchasing department. The issuance of the purchase requisition may be automated in the accounting information system so that it occurs automatically when the inventory reaches the reorder point.
- A **bill of materials** for each product is used to show the components needed and the quantities of each needed to manufacture a single unit of product.
- The **master production schedule** shows the quantities of goods needed to meet anticipated sales and when the quantities need to be produced in order to fulfill sales projections and maintain desired inventory levels.
- Labor time needs to be tracked for costing purposes. Job time cards may be used to capture the distribution of labor to specific orders.
- Enterprise Resource Planning systems are used in most large- and medium-sized firms to collect, store, manage, and interpret data across the organization. ERP systems can help a manufacturer to track, monitor, and manage production planning, raw materials purchasing, and inventory management, and are also integrated with tracking of sales and customer service.

Data entry is accomplished with automated technology such as bar code scanners, RFID (radio frequency identification systems), GPS locators, and other input technologies that can reduce input errors and support fast and accurate data collection for production.

RFID can be used to track components to products and the products themselves along the production process and the supply chain. Tags containing electronically-stored information are attached to components and products.

Example: An RFID tag attached to an automobile can track its progress through the assembly line, and RFID tags attached to individual components of the automobile can be used to make sure everything is assembled properly.

Outputs of the production cycle include:

- Materials price lists showing prices paid for raw materials, kept up to date by the purchasing department and used by cost accountants to determine actual and standard costs for production.
- Usage reports showing usage of raw materials by various production departments, used by cost accountants and managers to detect waste by comparing actual raw material usage to standard raw material usage for the actual production.
- Inventory reconciliations comparing physical inventories with book balances.
- Inventory status reports that enable purchasing and production managers to monitor inventory levels.
- Production cost reports detailing actual costs for cost elements, production processes, or jobs. These reports are used by cost accountants to calculate variances for materials, labor, and overhead.
- Manufacturing status reports, providing managers with information about the status of specific processes or jobs.
- Reports output from the production process are used in developing financial statements.

Human Resources and Payroll Cycle

The human resources management and payroll cycle involves hiring, training, paying, and terminating employees.

The accounting information system is used for:

- Recording the hiring and training of employees.
- Processes associated with employee terminations.
- Maintaining employee earnings records.
- Complying with regulatory reporting requirements, including payroll tax withholdings.
- Reporting on payroll benefit deductions such as for pensions or medical insurance.
- Making timely and accurate payroll payments to employees, payments for benefits such as pensions and medical insurance, and payroll tax payments to taxing authorities.

Inputs to the human resource management and payroll cycle include forms sent by the human resources department to payroll processing:

- Personnel action forms documenting hiring of employees and changes in employee pay rates and employee status.
- Time sheets or time cards that record hours worked.
- Payroll deduction authorization forms.
- Tax withholding forms.

Outputs of the human resource management and payroll cycle include:

- Employee listings showing current employees and information about them such as home addresses.
- Payroll payment registers listing gross pay, deductions, and net pay for each employee, used to make journal entries to the general ledger. The journal entries may be automated in the AIS.
- Preparing paychecks or, for direct deposits to employees' bank accounts, electronic funds transfers.
- Deduction reports containing company-wide or individual segment employee deduction information.
- Payroll summaries used by managers to analyze payroll expenses.
- Tax reports, used for remitting payroll taxes to the taxing authorities, both amounts withheld from employees' pay and employer taxes.
- Reporting to employees the information on their income and taxes paid that they need for their personal tax reporting.

The payroll process is outsourced by many companies.

Financing Cycle

The financing process is responsible for acquiring financial resources by borrowing cash or selling stock and for investing financial resources. It involves managing cash effectively, minimizing the cost of capital,[51] investing in a manner that balances risk and reward, and making cash flow projections in order to make any adjustments necessary to have cash on hand when it is needed for operations, investing, and repaying debt.

Minimizing the cost of capital requires determining how much capital should be in the form of debt and how much in the form of equity. **Financial planning models** are often used by management to help them determine the optimum strategies for acquiring and investing financial resources.

Managing cash effectively includes collecting cash as quickly as possible, and many firms use lockbox arrangements to reduce the collection time for payments received.[52] Managing cash effectively also means paying invoices as they are due and taking advantage of discounts for prompt payment when the discounts offered are favorable. Electronic funds transfer through the automated clearing house is often used to pay accounts payable and also to pay employees by depositing the funds directly to the payees' bank accounts. Use of electronic funds transfer enables a business to closely control the timing of funds being deducted from its operating and payroll accounts.

Projecting cash flows involves using a cash receipts forecast—an output of the revenue cycle—and cash disbursements forecasts—outputs of the purchasing and expenditures and the human resources and payroll cycles.

---

[51] "Capital" as used in the context of the "cost of capital" is the term used for the long-term funding used by firms that is supplied by its lenders or bondholders and its owners (its shareholders). A company's capital consists of its long-term debt and its equity. A company's cost of capital is the return expected by investors on a portfolio consisting of all the company's outstanding long-term debt and its equity securities. The cost of capital is tested on the CMA Part 2 exam and is covered in more depth in the study materials for that exam.

[52] With a lockbox system, a company maintains special post office boxes, called lockboxes, in different locations. Invoices sent to customers contain the address of the lockbox nearest to each customer as that customer's remittance address, so customers send their payments to the closest lockbox. The company then authorizes local banks with which it maintains deposit relationships to check these post office boxes as often as is reasonable, given the number of receipts expected. Because the banks are making the collections, the funds that have been received are immediately deposited into the company's accounts without first having to be processed by the company's accounting system, thereby speeding up cash collection. Cash management and the use of lockboxes are tested on the CMA Part 2 exam and are covered in more depth in the study materials for that exam.

Although the finance department utilizes the accounting information system, finance responsibilities should be segregated from accounting responsibilities. A common approach is to have a controller who manages the accounting function and a treasurer or CFO who manages the finance function.

The accounting information system is used for:

- The AIS can provide information about how quickly customers are paying their bills and can show trends in cash collections for use in managing the collection of cash.
- An AIS with EFT capability can be used to make payments by electronic funds transfer through the automated clearing house, or a separate EFT application that interfaces with the AIS can be used.
- Estimates of interest and dividend payments and receipts are used to develop cash flow forecasts.

Activities of and inputs to the financing cycle mostly originate outside the organization, as follows:

- Paper checks and remittance advices returned by customers with their payments are used to apply the payments to customers' accounts.
- Deposit receipts issued by the bank are used to document bank deposits.
- Bank statements are used to reconcile the cash balance according to the company's ledger with the cash balance in the bank account.
- Economic and market data, interest rate data and forecasts, and financial institution data are used in planning for financing.

Outputs of the financing cycle

The production of periodic financial statements draws on general ledger information about the financing processes, as follows:

- Interest revenue and expense.
- Dividend revenue and dividends paid.
- Summaries of cash collections and disbursements.
- Balances in investment accounts and in debt and equity.
- Cash budget showing projected cash flows.

Reports about investments and borrowings produced by the AIS for the financing cycle include:

- Changes in investments for a period.
- Dividends paid.
- Interest earned.
- New debt and retired debt for the period, including payments of principal and interest made for the period and information about lending institutions and interest rates.
- Significant ratios such as return on investment for the organization as a whole and for individual segments of it can be calculated by a financial planning model to help management make decisions regarding investing and borrowing.

Property, Plant, and Equipment (Fixed Asset) System

The fixed asset management system manages the purchase, valuation, maintenance, and disposal of the firm's fixed assets, also called property, plant, and equipment.

The accounting information system is used for:

- Recording newly-purchased fixed assets in the fixed asset module and the general ledger.
- Maintaining depreciation schedules and recording depreciation in order to calculate the book values of fixed assets.
- Tracking differences between depreciation as calculated for book purposes and depreciation as calculated for tax purposes in order to maintain deferred tax records.
- Maintaining records of the physical locations of fixed assets, as some of them may be moved frequently.
- Tracking repair costs and distinguishing between repair costs that are expensed and repair costs that are capitalized.
- Recording impairment of fixed assets.
- Tracking disposal of fixed assets and calculating the amount of gain or loss on the sale.

Activities and inputs to the fixed asset management system include:

- A request for a fixed asset purchase. The individual making the request uses a purchase requisition form, usually input electronically. The request usually requires approval by one or more higher-level managers.
- The purchasing department usually gets involved in vendor selection and issues a purchase order, similar to the way it is done for other purchases. Receiving reports and supplier invoices are handled the way they are for other purchases.
- If the company builds the fixed asset rather than acquiring it, a work order is used that details the costs of the construction.
- Repair and maintenance records need to be maintained for each fixed asset or for categories of fixed assets. Activities should be recorded on a repair and maintenance form so either the asset account can be updated or an expense account can be debited in the AIS, as appropriate, for the cost.
- When an existing fixed asset is moved from one location to another, those responsible should complete a fixed asset change form so that its location can be tracked.
- A fixed asset change form should also be used to record the sale, trade, or retirement of fixed assets.

Outputs of the fixed asset management system include:

- A list of all fixed assets acquired during a particular period.
- A fixed asset register listing the assigned identification numbers of each fixed asset held and each asset's location as of the register date.
- A depreciation register showing depreciation expense and accumulated depreciation for each fixed asset owned.
- Repair and maintenance reports showing the current period's repair and maintenance expenses and each fixed asset's repair and maintenance history.
- A report on retired assets showing the disposition of fixed assets during the current period.

All of the reports are used in developing information for the external financial statements.

The General Ledger and Reporting Systems

The general ledger contains accounts for all the assets, liabilities, equity, revenues, and expenses. The individual modules (or journals) support the general ledger and provide the detail behind the activity recorded in the general ledger.

In a responsibility accounting system, a journal such as accounts receivable can be subdivided according to responsibility center using responsibility center codes. Subsidiary ledgers are maintained for each responsibility center containing only the accounts used by that responsibility center.

In an automated accounting information system, most day-to-day transactions are initially recorded in the individual modules such as the accounts receivable or accounts payable module. Recording transactions in a module updates that module and possibly other modules and automatically creates transactions to the general ledger. Thus, the general ledger obtains data from the other cycles and processes it so that financial reports may be prepared. To be in accordance with generally accepted accounting principles, however, many valuation and adjusting entries are also required.

Financial Reporting Systems

The primary purpose of the financial reporting system is to produce external financial statements for the company's stakeholders and other external users such as analysts. The reports include the statement of financial position (balance sheet), income statement, statement of cash flows, statement of comprehensive income, and statement of changes in stockholders' equity.

Various internal reports are used by accountants who are reviewing data and making adjusting entries in order to produce the external financial statements. Two of them are:

- A **trial balance** is a columnar report that lists each account in the general ledger in the first column, followed by its balance as of a certain date and time in either the second or the third columns from the left. Debit balances are shown in the second column from the left and credit balances are shown in the third column from the left. All the debits are totaled and all the credits are totaled, and the total debits and total credits on the trial balance must balance. A trial balance is used to check preliminary balances before adjusting entries are made. It is printed after the adjustments are recorded to confirm that the adjusted balances are correct.
- A **general ledger report** is a report covering a specific period of time that shows either all the individual general ledger accounts and the transactions that adjusted them during that period, or it may be printed for only one or for only a few accounts. A general ledger report is used to analyze transactions posted to a specific account.

Management Reporting Systems

The information in the external financial statements is not what managers need to know for their decision making. Managers need the detailed internal statements produced by the AIS.

Cost accounting systems collect labor, material, and overhead costs that are used to determine the inventory costs of manufactured goods. Their output is used to determine the value of the inventories reported on the balance sheet, but cost accounting systems are also used to report variances from anticipated costs that production managers need in order to control production, as described in Section C, *Performance Management*, in Volume 1 of this textbook.

Profitability reporting systems and **responsibility reporting systems** involve comparisons between actual and planned amounts and variances. They are usually prepared according to responsibility center. Responsibility reporting traces events to the responsibility of a particular responsibility center or a particular manager. Each significant variance should be explained by the manager who has knowledge of what caused the variance. A variance may be a favorable variance, and an explanation of how it occurred may be useful information for other responsibility centers in the organization. If a variance is unfavorable, knowing how it occurred can be the first step to taking corrective action.

Databases

A database is a collection of related data files, combined in one location to eliminate redundancy, that can be used by different application programs and accessed by multiple users.

Basic Data Structure

The most commonly-used type of database is a **relational database**, which is a group of related tables. When the database is developed, specific data fields and records are defined. The data must be organized into a logical structure so it can be accessed and used. Data is stored according to a **data hierarchy**, and the data is structured in **levels**.

A data **field** is the first level in the data hierarchy. A field is information that describes one attribute of an item, or entity, in the database such as a person or an object. In an employee file, for example, one data field would be one employee's last name. Another field would be the same employee's first name. A field may also be called an "attribute," or a "column."

A database **record** is the second level of data. A database record contains all the information about one item, or entity, in the database. For example, a single database record would contain information about one employee. Each item of information, such as the employee's Employee ID number, last name, first name, address, department code, pay rate, and date of hire, is in a separate **field** within the employee's **record**. The data fields contained in each record are part of the **record structure**. The number of fields in each record and the size of each field is specified for each record.

A **file**, also called a **table**, is the third level of the data hierarchy. A table is a set of common records, such as records for all employees.

A complete **database** is the highest level. Several related **files** or **tables** make up a database. For example, in an accounting information system, the collection of tables will contain all the information needed for an accounting application.

Example: Consider the example of worker at a company who, over time, will have received numerous monthly paychecks. The **relational database** will contain at least two database **files**, or **tables**, for employees. The first table, the "Employees" table, contains all the **records** of the individual employees' IDs and their names. The second table, the "Paychecks" table, contains data on all the paychecks that have been issued and, for each paycheck, the Employee ID of the employee to whom it was issued.

A **database management system** can be used to locate all of the paychecks issued for one particular employee by using the employee ID attached to the person's name in that employee's **record** in the Employees table and locating all of the individual paycheck **records** for that same employee ID in the Paychecks table.

The Employee ID ties the information in the Paychecks table to the information in the Employees table.

Database Keys

The **primary key** is a data field in a record that distinguishes one record from another in the table. Every record in a database has a primary key, and each primary key is unique. The primary key is used to find a specific record, such as the record for a specific employee. A primary key may consist of one data field or more than one data field. For example, in an Employees table, each employee record contains an Employee ID. The Employee ID is the primary key in the Employees table.

Every record will have a primary key, and some records will also have **foreign keys**. Foreign keys connect the information in a record to one or more records in other tables.

Example: Using the example of employees, the first table, the "Employees" table, contains all the records of the individual employees' IDs and their names. The second table, the "Paychecks" table, contains records of all the paychecks that have been issued and, for each paycheck, the Employee ID of the employee to whom it was issued. In the Employees table, the Employee ID in each employee record serves as the **primary key**. In the Paychecks table, the Employee ID in each paycheck record is a **foreign key**.

Entity-Relationship Modeling

Database administrators use the Entity-Relationship Model to plan and analyze relational database files and records. An **entity-relationship diagram** utilizes symbols to represent the relationships between and among the different entities in the database. The three most important relationship types are **one-to-one**, **one-to-many**, and **many-to-many**. These relationship types are known as **database cardinalities** and show the **nature of the relationship** between the entities in the different files or tables within the database.

Example: Using the example of employees again, the connection between each employee (one person) and each employee's paychecks (which are many) is an example of a **one-to-many relationship**.

Database Management System (DBMS)

A database management system is a software package that serves as an **interface** between users and the database. A database management system manages a set of interrelated, centrally-coordinated data files by standardizing the storage, manipulation, and retrieval of data. It is used to create the database, maintain it, safeguard the data, and make the data available for applications and inquiries. Because of its standardized format, a database can be accessed and updated by multiple applications.

Database management systems perform four primary functions:

1) **Database development**. Database administrators use database management systems to develop databases and create database records.
2) **Database maintenance**. Database maintenance includes record deletion, alteration, and reorganization.
3) **Database interrogation**. Users can retrieve data from a database using the database management system and a **query language** in order to select subsets of records to extract information.
4) **Application development**. Application development involves developing queries, forms, reports, and labels for a business application and allowing many different application programs to easily access a single database.

Note: A database management system is not a database but rather a set of separate computer programs that enables the database administrator to create, modify, and utilize database information; it also enables applications and users to query the database.

A database management system provides **languages** for database development, database maintenance, and database interrogation, and it provides programming languages to be used for application development. The languages use **statements**, which is another word for **commands**. For example, a database administrator uses statements to create a database. A database administrator uses the DBMS not only to create a database, but also sometimes to create applications that will access the data in the database.

Database Development

When a relational database is developed, the data fields and records to be used in the database must be **structured**. The database administrator uses a database management system and a **Data Definition Language** (**DDL**) to create a description of the logical and physical structure or organization of the database and to structure the database by specifying and defining data fields, records, and files or tables. The database administrator also specifies how data is recorded, how fields relate to each other, and how data is viewed or reported.

The structure of the database includes the database's **schema**, **subschemas**, and **record structures**.

- The **schema** is a map or plan of the entire database—its logical structure. It specifies the names of the data elements contained in the database and their relationships to the other data elements.
- A particular application or user may be limited to accessing only a subset of the information in the database. The limited access for an application or a user is called a **subschema** or a **view**. One common use of views is to provide read-only access to data that anyone can query, but only some users can update. Subschemas are important in the design of a database because they determine what data each user has access to while protecting sensitive data from unauthorized access.

  Note: The schema describes the **design** of the database, while the subschemas describe the **uses** of the database. The database schema should be flexible enough to permit creation of all of the subschemas required by the users.

- In defining the **record structure** for each table, the database administrator gives each field a name and a description, determines how many characters the field will have, and what type of data each field will contain (for example, text, integer, decimal, date), and may specify other requirements such as how much disk space is needed.

The database administrator also defines the **format** of the input (for example, a U.S. telephone number will be formatted as [XXX] XXX-XXXX).

The **input mask** for a data field creates the appearance of the input screen a user will use to enter data into the table so that the user will see a blank field or fields in the style of the format. For example, a date field will appear as _____/_____/ .....The input mask helps ensure input accuracy.

Once the record structure of the database table is in place, the records can be created.

Database Maintenance

A **data manipulation language** (**DML**) is used to maintain a database and consists of "insert," "delete," and "update" statements (commands). Databases are usually updated by means of transaction processing programs that utilize the data manipulation language. As a result, users do not need to know the specific format of the data manipulation commands.

Database Interrogation

Users can retrieve data from a database by using a **query language**. **Structured Query Language** (**SQL**) is a query language, and it is also a data definition language and a data manipulation language. SQL has been adopted as a standard language by the American National Standards Institute (ANSI). All relational databases in use today allow the user to query the database directly using SQL commands. SQL uses the

"select" command to query a database. However, business application programs usually provide a **graphical user interface** (**GUI**) that creates the SQL commands to query the database for the user, so users do not need to know the specific format of SQL commands.

Application Development

Database management systems usually include one or more **programming languages** that can be used to develop custom applications by writing programs that contain statements calling on the DBMS to perform the necessary data handling functions. When writing a program that uses a database that is accessed with a DBMS, the programmer needs only the **name** of the data item, and the DBMS locates the data item in the storage media.

Note: One of the key characteristics of a database management system is that the **applications that access the database are programmed to be independent of the data itself**, meaning the programs do not refer to a specific number or item, but rather to the **name** of the data item.

This independence is similar to changing a number in a spreadsheet cell that is referenced in a formula in another cell elsewhere in the spreadsheet. It is not necessary to change the formula because the formula relates to the cell and not to the number itself. Whatever number appears in that cell will be used in the formula in the other cell.

Enterprise Resource Planning Systems

An accounting information system utilizes a database specific to the accounting, finance, and budgeting functions. Information systems are used throughout organizations for much more than financial applications, though, and they all require their own databases. For example, materials requirements planning systems are used to determine what raw materials to order for production, when to order them, and how much to order. Personnel resource systems are used to determine personnel needs and to track other personnel data.

However, problems arise when an organization's various systems do not "talk" to one another. For instance:

- Production is budgeted based on expected sales, and raw materials need to be ordered to support the budgeted production. If the materials requirements planning system cannot access the financial and budgeting system, someone needs to manually input the planned production of every product into the materials requirements planning system after the sales have been budgeted and the budgeted production has been set.
- A salesperson takes an order. If the items ordered are in stock, the salesperson submits the order to an order entry clerk, who prepares the invoice and shipping documents. The documents are delivered manually to the shipping department, and the shipping department prepares the shipment and ships the order. After shipping, the sale is recorded in the accounting information system and the customer's account is updated with the receivable due. The order information is entered separately into the database used by the customer relations management system, so that if the customer calls about the order, the customer service personnel will be able to locate the order information, because the customer service agents do not have access to the accounting records.

Inputting what is basically the same information multiple times is duplication of effort. Not only does it waste time, but the multiple input tasks cause delays in making the information available to those who need it. Furthermore, each time information is input manually into a separate system, the opportunity for input errors increases. Thus, when the information finally is available, it may not even be accurate.

Enterprise Resource Planning (ERP) can help to overcome the challenges of separate systems because it integrates all aspects of an organization's activities—operational as well as financial—into a single system that utilizes a single database.

ERP systems consist of the following components:

- Production planning, including determining what raw materials to order for production, when to order them, and how much to order.
- Logistics, both inbound (materials management) and outbound (distribution).
- Accounting and finance.
- Human resources.
- Sales, distribution, and order management.

Features of ERP systems include:

1) **Integration**. The ERP software integrates the accounting, customer relations management, business services, human resources, and supply chain management so that the data needed by all areas of the organization will be available for planning, manufacturing, order fulfillment, and other uses. The system tracks all of a firm's resources, including cash, raw materials, inventory, fixed assets, and human resources, forecasts their requirements, and tracks shipping, invoicing, and the status of commitments such as orders, purchase orders, and payroll.

2) **Centralized database**. The data from the separate areas of the organization flows into a secure and centralized database rather than several separate databases in different locations. All users use the same data that has been derived through common processes.

3) **Usually require business process reengineering**. An ERP system usually forces organizations to reengineer or redesign their business processes in order to use the system. Because ERP software is "off-the-shelf" software, customization is usually either impossible or prohibitively expensive. Thus, business processes used must accommodate the needs of the system and many may need to be redesigned.

When budgeted production is set, the information will immediately be available to determine what raw materials should be ordered, how much, and when. When a salesperson enters a customer's order into the system (or when it is automatically entered as a result of an online order), inventory availability can be immediately checked. If the order is a credit order, the customer's credit limit and payment history is checked. If everything is OK, the warehouse is notified to ship, the accounting information system is automatically updated, and if the order is a credit card order, the customer's credit card is charged. Information on the order and its status is immediately visible to customer service personnel so they can answer questions about the order if they receive an inquiry from the customer.

Information about exceptions is also immediately available and can be addressed automatically. For example:

- If the customer's credit card charge is declined, shipment of the order can be automatically held until an investigation can be performed and the order possibly cancelled.
- If the item ordered is not in stock, a backorder report is automatically generated for follow-up with the customer, and the ERP system can trigger the production system to manufacture more product. The production system can revise the production schedules accordingly, and human resources may be involved if additional employees will be required.

Extended ERP Systems

Extended enterprise resource planning systems include customers, suppliers, and other business partners. The systems interface with customers and suppliers through **supply chain management** applications that give partners along the supply chain access to internal information of their suppliers and customers. Suppliers can access the company's internal information such as inventory levels and sales orders, enabling the company to reduce its cycle time for procuring raw materials for manufacturing or goods for sale. Customers can also view their supplier's information about their pending orders.

Advantages of ERP Systems

- Integrated back-office systems result in better customer service and production and distribution efficiencies.
- Centralizing computing resources and IT staff reduces IT costs versus every department maintaining its own systems and IT staff.
- Centralization of data provides a secure location for all data that has been derived through common processes, and all users are using the same data.
- Day-to-day operations are facilitated. All employees can easily gain access to real-time information they need to do their jobs. Cross-functional information is quickly available to managers regarding business processes and performance, significantly improving their ability to make business decisions and control the factors of production. As a result, the business is able to adapt more easily to change and quickly take advantage of new business opportunities.
- Business processes can be monitored in new and different ways, such as with dashboards.
- Communication and coordination are improved across departments, leading to greater efficiencies in production, planning, and decision-making that can lead to lower production costs, lower marketing expenses, and other efficiencies.
- Data duplication is reduced and labor required to create inputs and distribute and use system outputs is reduced. Potential errors caused by inputting the same data multiple times are reduced.
- Expenses can be better managed and controlled.
- Inventory management is facilitated. Detailed inventory records are available, simplifying inventory transactions. Inventories can be managed more effectively to keep them at optimal levels.
- Trends can be more easily identified.
- The efficiency of financial reporting can be increased.
- Resource planning as a part of strategic planning is simplified. Senior management has access to the information it needs in order to do strategic planning.

Disadvantages of ERP Systems

- Business re-engineering (developing business-wide integrated processes for the new ERP system) is usually required to implement an ERP system and it is time-consuming and requires careful planning.
- Converting data from existing systems into the new ERP system can be time-consuming and costly and, if done incorrectly, can result in an ERP system that contains inaccurate information.
- Training employees to use the new system disrupts existing workflows and requires employees to learn new processes.
- An unsuccessful ERP transition can result in system-wide failures that disrupt production, inventory management, and sales, leading to huge financial losses. Customers who are inconvenienced by the implementation may leave. Because the entire business relies on the new ERP system, it is critical that it be completely functional and completely understood by all employees **before** it "goes live." No opportunities are available to "work out the bugs" or "learn the ropes" when the entire business relies on the one system.
- Ongoing costs after implementation include hardware costs, system maintenance costs, and upgrade costs.

Data Warehouse, Data Mart, and Data Lake

Data Warehouse

A copy of all of the historical data for the entire organization can be stored in a single location known as a **data warehouse**, or an **enterprise data warehouse**. A data warehouse is separate from an ERP system because a data warehouse is not used for everyday transaction processing. By having all of the company's information from different departments in one location for analysis, a company is able to more efficiently manage and access the information for data mining to discover patterns in the data for use in making decisions. For example, the marketing department can access production data and be better able to inform customers about the future availability of products.

Managers can use **business intelligence tools** to extract information from the data warehouse. For instance, a company can determine which of its customers are most profitable or can analyze buying trends.

Note: Business intelligence is the use of software and services to collect, store, and analyze data produced by a firm's business activities. Business intelligence tools are used to access and analyze data generated by a business and present easy-to-understand reports, summaries, dashboards, graphs, and charts containing performance measures and trends that provide users with detailed information about the business that can be used to make strategic and tactical management decisions. Business intelligence is covered in more detail later in this section.

To be useful, data stored in a data warehouse should:

1) Be free of errors.
2) Be uniformly defined.
3) Cover a longer time span than the company's transactions systems to enable historical research.
4) Allow users to write queries that can draw information from several different areas of the database.

Note: The data in a data warehouse is a **copy** of historical data, and therefore is not complete with the latest real-time data. Furthermore, information in a data warehouse is read-only, meaning users cannot change the data in the warehouse.

Because the data stored in a data warehouse exists in different formats in the various sources from which it is copied, all differences need to be resolved to make the data available in a unified format for analysis. The process of making the data available in the data warehouse involves the following.

1) Periodically, data is uploaded from the various data sources, usually to a staging server before going to the data warehouse. The data upload may occur daily, weekly, or with any other established frequency.
2) The datasets from the various sources are transformed to be compatible with one another by adjusting formats and resolving conflicts. The transformation that must take place before the data can be loaded into a data warehouse is known as **Schema-on-Write** because the schema is applied before the data is loaded into the data warehouse.
3) The transformed data is loaded into the data warehouse to be used for research, analysis, and other business intelligence functions.

Data Mart

A **data mart** is a subsection of a data warehouse that provides users with analytical capabilities for a restricted set of data. For example, a data mart can provide users in a department such as accounts receivable access to only the data that is relevant to them so that the accounts receivable staff do not need to sift through unneeded data to find what they need.

A data mart can provide security for sensitive data because it isolates the data certain people are authorized to use and prevents them from seeing data that needs to be kept confidential. Furthermore, because each data mart is used only by one department, the demands on the data servers can be distributed; one department's usage does not affect other departments' workloads.

Data marts can be of three different types:

1) A **dependent** data mart draws on an existing data warehouse. It is constructed using a top-down approach and withdraws a defined portion of the data in the data warehouse when it is needed for analysis. Several dependent data marts, used by different areas of the organization, can draw on a single data warehouse.

2) An **independent** data mart is created without the use of a data warehouse through a bottom-up approach. The data for just one data mart for a single business function is uploaded, transformed, and then loaded directly into the data mart. If an organization needs several data marts for different areas of the organization, each one would need to be created and updated separately, which is not optimal. Furthermore, the idea of independent data marts is antithetical to the motivation for developing a data warehouse.

3) A **hybrid** data mart combines elements of dependent and independent data marts, drawing some data from an existing data warehouse and some data from transactional systems. Only a hybrid data mart allows analysis of data from a data warehouse with data from other sources.

Data Lake

Much of the data captured by businesses is **unstructured**, such as social media data, videos, emails, chat logs, and images of invoices, checks, and other items. Such data cannot be stored in a data warehouse because the types of data are so disparate and unpredictable that the data cannot be transformed to be compatible with the data in a data warehouse. A **data lake** is used for unstructured data.

A data lake is a massive body of information fed by multiple sources for which the content has not been processed. Unlike data warehouses and data marts, data lakes are not "user friendly." Data lakes have important capabilities for data mining and generating insights, but usually only a data scientist is able to access it because of the analytical skills needed to make sense of the raw information.

A data lake utilizes a non-relational database management system, called **NoSQL**, which stands for "Not only SQL." A NoSQL database management system can be used to analyze high volume and disparate data, including unstructured and unpredictable data types. Unlike relational databases, NoSQL databases do not require SQL to analyze the data and most do not use a schema and thus they are more flexible.[53] A NoSQL database management system can be used with a data lake that contains both structured and unstructured data.

Note: SQL can be used as a query language with a NoSQL database management system, but SQL is not the main query language used because its usage is limited to structured data.

[53] A database's **schema** is a map or plan of the entire database, that is, the database's logical structure. The schema specifies the names of the data elements contained in the database and their relationships to the other data elements. For more information about relational databases, please see the topic *Databases* in this volume in Section F.1. – *Information Systems* in this volume.

Enterprise Performance Management

Enterprise Performance Management (EPM), also known as Corporate Performance Management (CPM) or Business Performance Management (BPM), is a method of monitoring and managing the performance of an organization in reaching its performance goals. It is the process of linking strategies to plans and execution.

The organization's strategic goals and objectives must be clearly communicated to managers and be incorporated into their budgets and plans. Then, periodically the performance of the organization is reviewed with respect to its progress in attaining the strategic goals and objectives. Key Performance Indicators (KPIs), Balanced Scorecards, and Strategy Maps are frequently used and are monitored and managed. If the organization or a segment of it is not performing as planned, adjustments are made, either in the strategy or in the operations.

Enterprise Performance Management software is available that integrates with an organization's accounting information system, ERP system, customer relations management system, data warehouse, and other systems. It is designed to gather data from multiple sources and consolidate it to support performance management by automating the collection and management of the data needed to monitor the organization's performance in relation to its strategy. Users can create reports and monitor performance using the data captured and generated in the other systems. Some examples of an EPM's capabilities include:

- Reports comparing actual performance to goals.
- Reports on attainment of KPIs by department.
- Balanced scorecards, strategy maps, and other management tools.
- Creating and revising forecasts and performing modeling.
- Generating dashboards presenting current information customized to the needs of individual users.

EPM software can also automate budgeting and consolidations. Tasks that in the past may have required days or weeks can now be completed very quickly.

Note: EPM software can be on premises or it can be deployed as Software as a Service (SaaS), otherwise known as "the cloud." Cloud computing is covered in this section in the topic *Technology-enabled Finance Transformation*.

Data Governance

Definition of Data Governance

Corporate governance includes all of the means by which businesses are directed and controlled, including the rules, regulations, processes, customs, policies, procedures, institutions, and laws that affect the way the business is administered. Corporate governance spells out the rules and procedures to be followed in making decisions for the corporation.

Data governance is similar, but it is specific to data and information technology. Data governance encompasses the practices, procedures, processes, methods, technologies, and activities that deal with the overall management of the data assets and data flows within an organization. Data governance is a process that helps the organization better manage and control its data assets. In a sense, data governance is quality control for data. It enables reliable and consistent data, which in turn makes it possible for management to properly assess the organization's performance and make management decisions.

Data governance includes the management of the following.

- **Data availability**, or the process of making the data available to users and applications when it is needed and where it is needed.
- **Data usability**, including its accessibility to users and applications, its quality, and its accuracy.
- **Data integrity**, or the completeness, consistency, reliability, and accuracy of data.
- **Data security**, meaning data protection, including prevention of unauthorized access and protection from corruption and other loss, including backup procedures.
- **Data privacy**, that is, determining who is authorized to access data and which items of data each authorized person can access.
- **Data integration**, which involves combining data from different sources (which can be both internal and external) and providing users with a unified view of all the data.
- **System availability**, that is, maximizing the probability that the system will function as required and when required.
- **System maintenance**, including modifications of the system done to correct a problem, to improve the system's performance, to update it, or to adapt it to changed requirements or a changed environment.
- **Compliance with regulations**, such as laws regulating privacy protections.
- Determination of **roles and responsibilities** of managers and employees.
- Internal and external **data flows** within the organization.

IT Governance and Control Frameworks

IT governance and control frameworks have been developed to provide **models**, or sets of standardized guidelines, for the management of IT resources and processes. Frameworks provide numerous benefits to an organization.

- They **identify specific roles** and responsibilities that need to be met.
- They **provide a benchmark** for assessing risks and controls.
- Following a framework provides a **higher likelihood of implementing effective governance and controls**.
- Frameworks **break down objectives and activities** into groups.
- **Regulatory compliance may be easier to achieve** by following effective governance and control frameworks.

Following is an overview of two of the most prominent IT governance frameworks currently in use: COSO's internal control framework and ISACA's COBIT.

Internal Control – Integrated Framework by COSO, the Committee of Sponsoring Organizations

The **Committee of Sponsoring Organizations** (**COSO**)[54], which consists of five professional organizations, created one of the first internal control frameworks in 1992 with the publication of *Internal Control—Integrated Framework* and it introduced the concept of controls for IT systems. *Internal Control—Integrated Framework* was updated in 2013. *Internal Control—Integrated Framework* is covered in this volume in *Section E, Internal Control*, so it will be reviewed only briefly here.

Internal Control—Integrated Framework defines internal control as "a process, effected by[55] an entity's board of directors, management, and other personnel, designed to provide reasonable assurance regarding the achievement of objectives relating to operations, reporting, and compliance." According to the *Integrated Framework*, the internal control system should consist of the following five interrelated components.

1) The **control environment**: the standards, processes, and structures that provide the foundation for carrying out internal control.

2) **Risk assessment**: the process of identifying, analyzing, and managing the risks that have the potential to prevent the organization from achieving its objectives, relative to the organization's established risk tolerance.

3) **Control activities**: the actions established by policies and procedures that help ensure that management's instructions intended to limit risks to the achievement of the organization's objectives are carried out.

4) **Information and communication**: obtaining, generating, using, and communicating relevant, quality information necessary to support the functioning of internal control. Communication needs to be both internal and external.

5) **Monitoring**: overseeing the entire internal control system to assess the operation of existing internal controls to ensure that the internal control system continues to operate effectively.

COBIT® by ISACA

COBIT® is an I & T (Information and Technology) framework for the governance and management of enterprise information and technology. "Enterprise information and technology" refers to all the technology and information processing used by the whole enterprise to achieve its goals, no matter where the technology and information processing occurs in the enterprise. Thus, while enterprise I & T includes the organization's IT department, it is not limited to the IT department.

The *COBIT*® Framework was first introduced for information systems in 1996 by ISACA, an independent, nonprofit, global association dedicated to the development, adoption, and use of globally-accepted knowledge and practices for information systems. ISACA was previously known as the Information Systems Audit and Control Association. However, ISACA now serves a wide constituency of other IT professionals, including consultants, educators, security professionals, risk professionals, and chief information officers. To reflect the fact that it now serves such a broad range of IT professionals, the organization is now known simply by its acronym, ISACA.

---

[54] The Committee of Sponsoring Organizations sponsored the Treadway Commission, the National Commission on Fraudulent Financial Reporting, that was created in 1985 to identify the causal factors of fraudulent financial reporting and to make recommendations to reduce its incidence. The sponsoring organizations included the AICPA (American Institute of Certified Public Accountants), the AAA (American Accounting Association), the IIA (Institute of Internal Auditors), the FEI (Financial Executives International), and the IMA (Institute of Management Accountants).

[55] To "effect" something means to cause it to happen, put it into effect, or to accomplish it. So "effected by" means "put into effect by" or "accomplished by."

In line with the early identity of the organization, the early focus of *COBIT*® was on information systems audit and control, and *COBIT*® was an acronym for **C**ontrol **OB**jectives for **I**nformation and Related **T**echnology. However, in the intervening years the focus of the Framework has changed to IT governance and management in recognition of the needs of the wide range of IT professionals that ISACA serves. When *COBIT*® *5* was introduced in 2012, ISACA dropped the Framework's full name entirely, and like ISACA, *COBIT*® is now known simply by its acronym.[56]

ISACA published an updated version of the Framework, *COBIT*® *2019*, in November 2018. The information that follows is from *COBIT*® *2019*.

COBIT® *2019* draws a distinction between governance and management. According to the Framework, governance and management are different disciplines that involve different activities, different organizational structures, and serve different purposes.

Governance is usually the responsibility of the board of directors under the leadership of the chair of the board of directors. The purpose of governance is to ensure that:

- Stakeholder[57] needs are considered and conditions and options are evaluated in order to determine balanced, agreed-on enterprise objectives.
- Prioritization and decision-making are used to set direction.
- Performance and compliance are monitored in terms of the agreed-on direction and enterprise objectives.

Management is usually the responsibility of the executive management under the chief executive officer's (CEO's) leadership. The purpose of management is to plan, build, run, and monitor activities, in accordance with the direction set by the body responsible for governance such as the board of directors, in order to achieve the enterprise objectives.[58]

The guidance in the *COBIT* Framework is generic in nature so that users can customize it to focused guidance for the enterprise.[59]

Components of a Governance System

Each enterprise needs to establish and sustain a governance system that includes the following components, or factors, that contribute to the operations of the enterprise's governance system over information and technology (I & T).

- **Processes**. A process is a set of practices and activities needed to achieve a specific objective and produce outputs that support achievement of IT-related goals.
- **Organizational structures**. Organizational structures are the primary decision-making entities within the enterprise.
- **Principles**, **policies**, **and frameworks**. Principles, policies, and frameworks provide practical guidance for the day-to-day management of the enterprise.
- **Information**. Information includes all the information produced and used by the enterprise. *COBIT*® *2019* focuses on the information needed for effective governance of the enterprise.

---

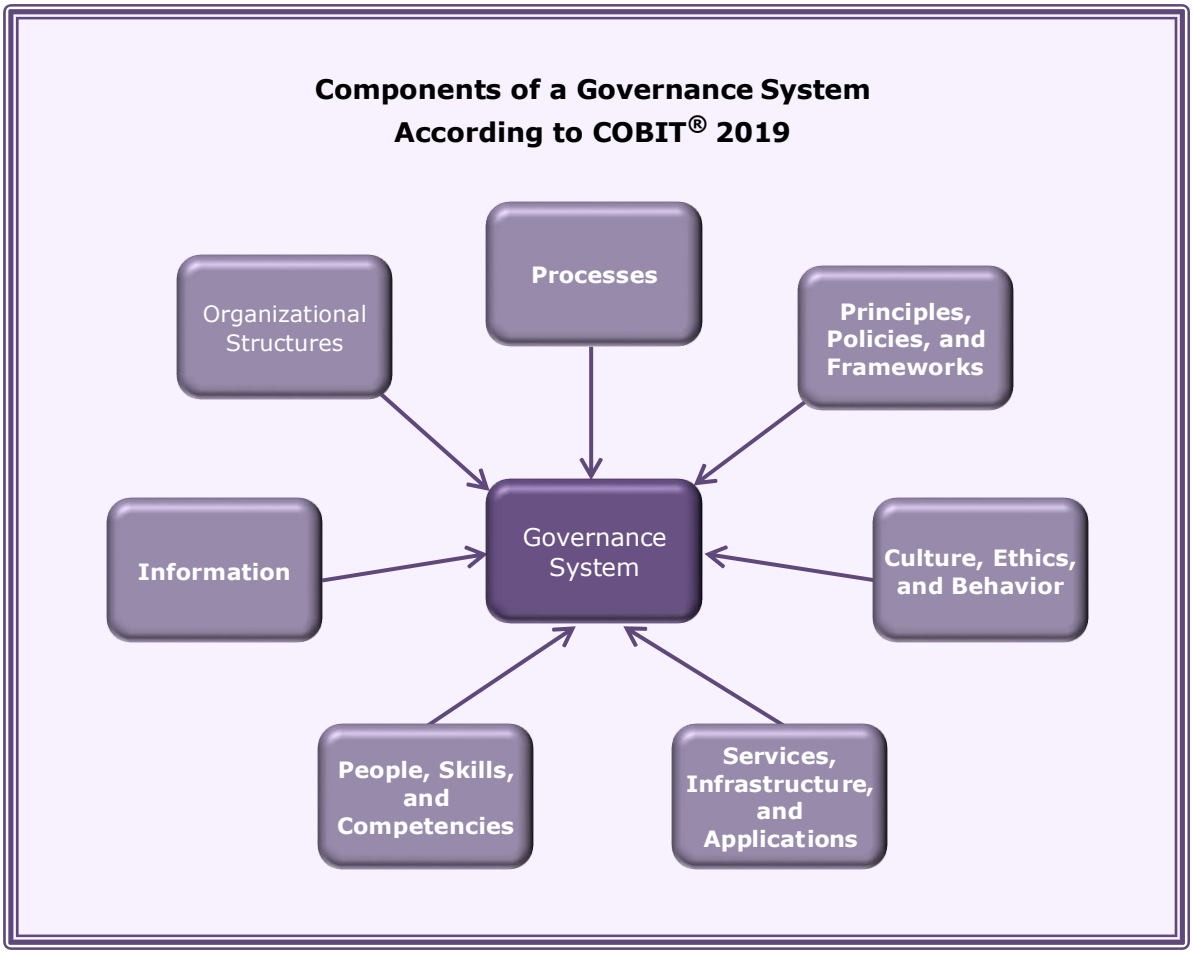
[56] ISACA also promulgates separate IS auditing and IS control standards, so IS audit and control have not been left behind.

[57] Stakeholders for enterprise governance of information and technology include members of the board of directors, executive management, business managers, IT managers, assurance providers such as auditors, regulators, business partners, and IT vendors. (*COBIT*® *2019 Framework: Introduction and Methodology*, p. 15, © 2018 ISACA. All rights reserved. Used by permission.) Stakeholders are discussed in more detail later in this topic.

[58] *COBIT*® *2019 Framework: Introduction and Methodology*, p. 13, © 2018 ISACA. All rights reserved. Used by permission.

[59] Ibid., p. 15.

- **Culture**, **ethics**, **and behavior**. The culture of the enterprise and the ethics and behavior of both the enterprise and the individuals in it are important factors in the success of governance and management activities.
- **People**, **skills**, **and competencies**. People and their skills and competencies are important for making good decisions, for corrective action, and for successful completion of activities.
- **Services**, **infrastructure**, **and applications**. These include the infrastructure, technology, and applications used to provide the governance system for information and technology processing.[60]



[60] Ibid., pp. 21-22.

Goals Cascade

Enterprise goals, one of the design factors for a governance system, are involved in transforming stakeholder needs into actionable strategy for the enterprise.

Stakeholders for enterprise governance of information and technology (EGIT) and the target audience for *COBIT*® include the following:

Internal stakeholders:

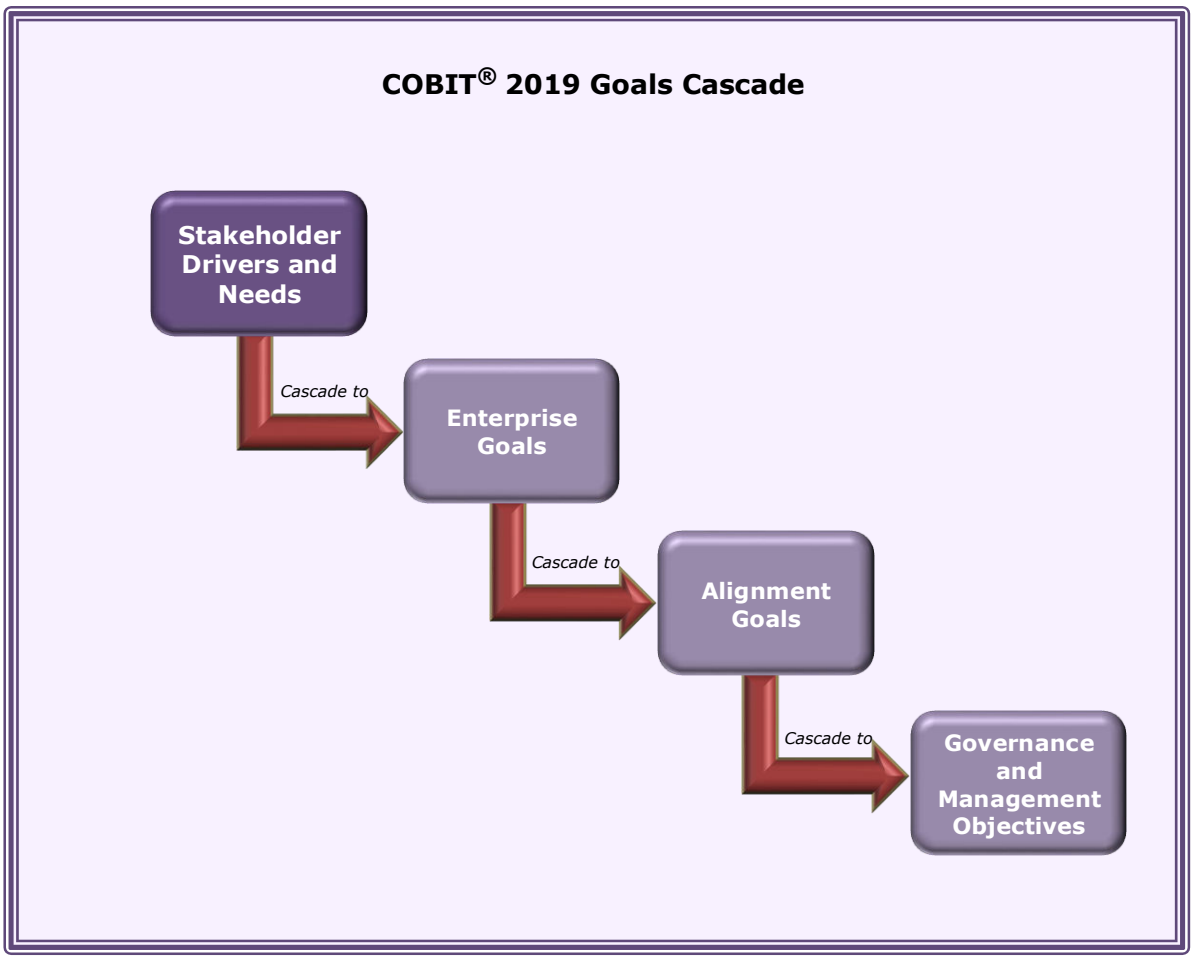
- **Members of the board of directors**, for whom *COBIT*® provides insight into how to obtain value from the use of I & T and explains relevant board responsibilities.
- **Executive management**, for whom *COBIT*® provides guidance in organizing and monitoring performance of I & T.
- **Business managers**, for whom *COBIT*® helps understanding in how to obtain the I & T solutions that the enterprise requires and how best to use new technology and exploit it for new strategic opportunities.
- **IT managers**, for whom *COBIT*® provides guidance in how best to structure and operate the IT department and manage its performance.
- **Assurance providers such as auditors**, for whom *COBIT*® helps in managing assurance over IT, managing dependency on external service providers, and ensuring an effective and efficient system of internal controls.
- **Risk management**, for whom *COBIT*® helps with identification and management of IT-related risk.

External stakeholders:

- **Regulators**, in relation to which *COBIT*® helps ensure an enterprise is compliant with applicable rules and regulations and has an adequate governance system in place to manage compliance.
- **Business partners**, in relation to which *COBIT*® helps the enterprise to ensure that a business partner's operations are secure, reliable, and compliant with applicable rules and regulations.
- **IT vendors**, in relation to which *COBIT*® helps the enterprise to ensure that IT vendors' operations are secure, reliable, and compliant with applicable rules and regulations.[61]

[61] Ibid., p. 15.

Management objectives are prioritized based on prioritization of enterprise goals, which in turn are prioritized based on stakeholder drivers and needs. Alignment goals emphasize the alignment of the IT efforts with the goals of the enterprise.[62]



[62] Ibid., p. 28.

Performance Management in *COBIT® 2019*

Performance management includes the activities and methods used to express how well the governance and management systems and the components of an enterprise work, and if they are not achieving the required level, how they can be improved. Performance management utilizes the concepts **capability levels** and **maturity levels**.[63]

Performance management is organized in *COBIT® 2019* according to the components that make up the enterprise's governance system over information and technology. Although not all of the components are specifically addressed in *COBIT® 2019* as to performance management issues at this time, to review, the components include:

- Processes
- Organizational structures
- Principles, policies and frameworks
- Information
- Culture, ethics, and behavior
- People, skills, and competencies
- Services, infrastructure, and applications.

Those components for which performance management issues have been addressed include the following.

Managing the Performance of the Processes Component of the I & T Governance System

Governance and management objectives consist of several processes, and a **capability level** (or **maturity level**) is assigned to all process activities. The capability level is an expression of how well the process is implemented and is performing. A process reaches a certain capability level when all the activities of that level are performed successfully. Capability levels range from 0 (zero) to 5, as follows:

Level	General Characteristics
0	Lack of any basic capability; incomplete approach to address governance and management purpose; may or may not be meeting the intent of any Process practices.
1	The process more or less achieves its purpose through the application of an incomplete set of activities that can be characterized as initial or intuitive—not very organized.
2	The process achieves its purpose through the application of a basic, yet complete, set of activities that can be characterized as performed.
3	The process achieves its purpose in a much more organized way using organizational assets. Processes typically are well defined.
4	The process achieves its purpose, is well defined, and its performance is (quantitatively) measured.
5	The process achieves its purpose, is well defined, its performance is measured to improve performance, and continuous improvement is pursued.[64]

[63] Ibid., p. 37.
[64] *COBIT® 2019 Framework: Governance and Management Objectives*, pp. 19-20, © 2018 ISACA. All rights reserved. Used by permission.

Data Life Cycle

The **data life cycle** encompasses the period from creation of data and its initial storage through the time the data becomes out of date or no longer needed and is purged. The stages of the data life cycle include data capture, data maintenance, data synthesis, data usage, data analytics, data publication, data archival, and data purging.

The stages do not describe sequential data flows, because data may pass through various stages several times during its life cycle. Furthermore, data does not have to pass through all of the stages. However, data governance challenges exist in all of the stages and each stage has distinct governance needs, so it is helpful to recognize the various stages and some of the governance challenges associated with each.

- **Data capture** is the process of creating new data values that have not existed before within the organization. Data can be captured through external acquisition, data entry, or signal reception.
    - **External acquisition**. Data can be acquired from an outside organization, often through contracts governing how the data may be used. Monitoring performance with the contracts is a significant governance challenge.
    - **Data entry**. Data can be created through entry by human operators or by devices that generate data. Data governance requires monitoring the accuracy of the input.
    - **Signal reception**. Data may be received by transmission, for example from sensors. Data governance challenges include monitoring the function of the devices and the accuracy of the data received.
- **Data maintenance** is the processing of data before deriving any value from it, such as performing data integration. A governance issue is determining how best to supply the data to the stages at which data synthesis and data usage occur.
- **Data synthesis** is the creation of new data values using other data as input. To "synthesize" something means to combine different things to make something new. Data synthesis is therefore combining data from different sources to create new data. Governance issues include concerns about data ownership and the need for citation, the quality and adequacy of the input data used, and the validity of the synthesized data.
- **Data usage**, which is the application of the data to tasks, whether used in support of the organization or used by others as part of a product or service the organization offers. A governance challenge with respect to data usage is whether the data can legally be used in the ways the users want to use it. Regulatory or contractual constraints on the use of the data may exist, and the organization must ensure that all constraints are observed.
- **Data analytics**, or the process of gathering and analyzing data in a way that produces meaningful information to aid in decision-making. As businesses become more technologically sophisticated, their capacity to collect data increases. However, the stockpiling of data is meaningless without a method of efficiently collecting, analyzing, and utilizing it for the benefit of the company. A governance issue with respect to data analytics is ensuring that the company's data is accurately recorded, stored, evaluated, and reported.
- **Data publication** occurs when data is sent outside of the organization or leaves it in other ways. Periodic account statements sent to customers are an example of data publication, but data breaches also constitute data publication. The governance issue is that once data has been released it cannot be recalled, because it is beyond the reach of the organization. If published data is incorrect or if a data breach has occurred, data governance is needed in deciding how to deal with it.
- **Data archival** is the removal of data from active environments and its storage in case it is needed again. No maintenance, usage, or publication occurs while the data is archived, but if it is needed, it can be restored to an environment where it will again be maintained, used, or published.

- **Data purging** occurs at the end of the data's life cycle. Data purging is the removal of every copy of a data item from all locations within the organization. Purging should generally be done only of data that has been previously archived. Data governance considerations include development and maintenance of data retention and destruction policies that comply with all laws and regulations regarding record retention; conformance with established policies; confirmation that purging has been done properly; and documentation of data purged.

Records Management

Every organization should have a documented records management policy that establishes how records are to be maintained, identified, retrieved, preserved, and when and how they are to be destroyed. The policy should apply to everything defined by the organization as a "record," which includes both paper documents and data records. The concern for information systems is, of course, data records.

Although not every item of data will be determined to be a "record," consideration in the policy should also be given to management of data that is not actually "records," such as drafts of documents. The records management policy should identify the information that is considered records and the information that is not considered records but that nevertheless should be subject to the guidance in the policy.

Factors to be considered in developing a records management policy include:

- **Federal**, **state**, **and local document retention requirements**. U.S. federal requirements include Internal Revenue Service requirements for retaining income tax information and various employment laws and laws governing employee benefits. State and local requirements may also apply. Regulations and laws provide minimum records retention requirements, but those should not be regarded as guidance on when records must be destroyed. Decisions may be made to retain specific records longer than their required minimum periods due to other factors such as ongoing business use or internal audit requirements.
- **Requirements of the Sarbanes-Oxley Act of 2002**. Section 802 of the Sarbanes-Oxley Act prohibits altering, destroying, mutilating, concealing, or falsifying records, documents, or tangible objects with the intent to obstruct, impede, or influence a potential or actual federal investigation or bankruptcy proceeding. Violation is punishable by fines and/or imprisonment for up to 20 years. Furthermore, accountants must maintain certain audit records or review work papers for a period of five years from the end of the fiscal period during which the audit or review was concluded. Section 1102 of the Act states that corruptly altering or destroying a record or other object with the intent to impair its integrity or availability for use in an official proceeding is also punishable with fines and/or imprisonment for up to 20 years.
- **Statute of limitations information**. A statute of limitations is the time period during which an organization may sue or be sued or the time period within which a government agency can conduct an examination.
- **Accessibility**. An important consideration with electronic records is hardware, software, and media obsolescence. Records can become inaccessible if they are in an obsolete format, and the records management policy must include a means to either migrate the records to new versions, or the old hardware and software must be retained so the records can be accessed. If the records are to be migrated to new formats, quality control procedures must be in place to ensure none of the content is lost or corrupted.
- **Records of records**. The records management policy should establish the practice of maintaining an index of active and inactive records and their locations and of maintaining logs containing records of all purged data.

Needless to say, it is not enough to simply have a documented records management policy. The policy must be followed consistently. However, if a lawsuit is pending or anticipated, all document and data destruction should cease. All documents and data should be preserved, even if they would otherwise be purged under the policy.

Electronic data management should be supported by top management, and responsibility for custody of the records should be assigned.

Benefits of a Documented and Well-Executed Records Management Policy

- Locating documents that are needed is easier and more efficient.
- In the event of an examination, investigation, or lawsuit, having a documented records management policy **that is consistently followed** demonstrates a legitimate and neutral purpose for having previously destroyed any requested documents or data purged in accordance with the policy.
- Having a documented policy **that is consistently followed** increases the probability of the organization's compliance with all federal, state, and local regulations relating to document and data retention and destruction.
- Records will be adequately protected and their accessibility maintained.
- Records that are no longer needed or that are no longer of value will be destroyed at the appropriate time.

Cyberattacks

Cybersecurity is the process or methods of protecting Internet-connected networks, devices, or data from attacks. Cyberattacks are usually made to access, change, or destroy data, interrupt normal business operations, or, as with ransomware, they may involve extortion.

Some specific cybersecurity risks include the following:

- **Copyright infringement** is the theft and replication of copyrighted material, whether intellectual property, such as computer programs or textbooks, or entertainment property such as music and movies.
- **Denial of Service** (**DOS**) attacks occur when a website or server is accessed so frequently that legitimate users cannot connect to it. **Distributed Denial of Service** (**DDOS**) attacks use multiple systems in multiple locations to attack one site or server, which makes stopping or blocking the attack difficult. Hackers gain access to unsecured **Internet of Things** (**IoT**)[65] devices on which the default passwords have not been changed and use malware tools to create a **botnet** made up of innumerable devices. A botnet is a network of devices connected through the Internet that are all infected with the malware. A hacker or a group of hackers control the botnet without the owners' knowledge. The hacker directs the botnet to continuously and simultaneously send junk Internet traffic to a targeted server, making it unreachable by legitimate traffic. Sophisticated firewalls and network monitoring software can help to mitigate DOS and DDOS attacks.
- **Buffer overflow attacks** are designed to send more data than expected to a computer system, causing the system to crash, permitting the attacker to run malicious code, or even allowing for a complete takeover of the system. Buffer overflow attacks can be easily prevented by the software programs adequately checking the amount of data received, but this common preventative measure is often overlooked during software development.

[65] Internet of Things (IoT) devices are products used in homes and businesses that can be controlled over the Internet by the owner. Examples are door locks, appliances, lights, energy- and resource-saving devices and other devices that can be controlled either remotely or on-premises using voice commands.

- **Password attacks** are attempts to break into a system by guessing passwords. **Brute force attacks** use programs that repeatedly attempt to log in with common and/or random passwords, although most modern systems effectively prevent brute force attacks by blocking login attempts after several incorrect tries. Two-factor authentication can also prevent brute force attacks from being successful because a password alone will not allow access to the system. Systems should include sophisticated logging and **intrusion-detection systems** to prevent password attacks, and password requirements should be in place to reject short or basic passwords such as "password" or "123456."
- **Phishing** is a high-tech scam that uses spam email to deceive people into disclosing sensitive personal information such as credit card numbers, bank account information, Social Security numbers, or passwords. Sophisticated phishing scams can mock up emails to look like the information request is coming from a trusted source, such as state or local government, a bank, or even a coworker. The best defense against **phishing** is awareness and common sense. Recipients should be wary about any email that requests personal or financial information and should resist the impulse to click on an embedded link.
- **Malware** broadly refers to malicious software, including viruses. **Spyware** can secretly gather data, such as recording keystrokes in order to harvest banking details, credit card information, and passwords. Other types of malware can turn a PC into a **bot** or **zombie**, giving hackers full control over the machine without alerting the owner to the problem. Hackers can then set up "botnets," which are networks consisting of thousands or millions of "zombies," which can be made to send out spam emails, emails infected with viruses, or as described above, to cause distributed denial of service attacks.
- **Ransomware** is particularly dangerous malware that encrypts data on a system and then demands a ransom (a payment) for decryption. If the ransom is not paid, the data is lost forever. The most common way that ransomware is installed is through a malicious attachment or a download that appears to come from a trusted source. The primary defenses against ransomware are to avoid installing it in the first place and having data backups.
- "**Pay-per-click**" **abuse** refers to fraudulent clicks on paid online search ads (for example, on Google or Bing) that drive up the target company's advertising costs. Furthermore, if there is a set limit on daily spending, the ads are pushed off the search engine site after the maximum-clicks threshold is reached, resulting in lost business as well as inflated advertising costs. Such scams are usually run by one company against a competitor.

Some cybercrime is conducted on a more personal and in-person fashion. Through **social engineering** an individual may pose as a trustworthy coworker, perhaps someone from the company's IT support division, and politely ask for passwords or other confidential information. **Dumpster diving** is the act of sifting through a company's trash for information that can be used either to break into its computers directly or to assist in social engineering.

Outsiders are not the only threat to the security of a company's systems and data. Insiders can also be a source of security risks. For example, disgruntled employees or those who are planning to work for a competitor can sabotage computer systems or steal proprietary information.

Defenses Against Cyberattack

Encryption is an essential protection against hacking. Encryption protects both stored data and data that could be intercepted during transmission. If a hacker gains access to encrypted files, the hacker is not able to read the information.

Ethical hackers are network and computer experts with hacking skills who attempt to attack a secured system. They use the same methods as are used by malicious hackers, but if they find vulnerabilities that could be exploited by a malicious hacker, they report them to the owner or manager so they can be remedied. Ethical hacking is called **intrusion testing**, **penetration testing**, and **vulnerability testing**.

Advanced firewalls are firewalls that perform traditional firewall protection but have other capabilities, as well. Traditional firewalls use **packet filtering** to control network access by monitoring outgoing and incoming packets. **Packets** are used in networking to carry data during transmission from its source to its destination. They have a header that contains information that is used to help them find their way and to reassemble the data after transmission. Traditional firewalls permit or deny access based on the information in the header about the source and the destination Internet Protocol (IP) addresses, protocols, and ports.

Advanced firewalls are called **Next Generation Firewalls** (NGFW). In addition to the traditional firewall protection, advanced firewalls can filter packets based on applications and can distinguish between safe applications and unwanted applications because they base their detection on packet contents rather than on information in packet headers. Thus, they are able to block malware from entering a network, which is something that traditional firewalls cannot do. However, both traditional and advanced firewalls rely on hardware. For cloud-based (Software as a Service, or SaaS) applications, a cloud-based (Firewall as a Service) application is needed.

Access Controls

Access controls provide additional defenses against cyberattack. Access controls include **logical access controls** and **physical access controls**.

Logical Access Controls

Companies need to have strict controls over access to their proprietary data. Poor data oversight can leave a company vulnerable to accidents, fraud, and malicious parties who manipulate equipment and assets. **Logical security** focuses on **who can use which computer equipment** and **who can access data**. **Logical access controls** identify authorized users and control the actions that they can perform.

To restrict data access only to authorized users, one or more of the following strategies can be adopted:

1) **Something you know**
2) **Something you are**
3) **Something you have**

Something You Know

User IDs and passwords are the most common "something you know" way of authenticating users. Security software can be used to encrypt passwords, require changing passwords after a certain period of time, and require passwords to conform to a certain structure (for example, minimal length, no dictionary words, restricting the use of symbols). Procedures should be established for issuing, suspending, and closing user accounts, and access rights should be reviewed periodically.

Something You Are

Biometrics is the most common form of "something you are" authentication. Biometrics can recognize physical characteristics such as:

- Iris or retina of the eyes
- Fingerprints
- Vein patterns
- Faces
- Voices

Biometric scanners can be expensive and are generally used only when a high level of security is required.

Something You Have

Some very high-security systems require the presence of a physical object to certify an authorized user's identity. The most common example of this "something you have" authentication is a fob, a tiny electronic device that generates a unique code to permit access; for increased security, the code changes at regular intervals. A lost fob may be inconvenient but not a significant problem because the fob by itself is useless. Furthermore, a stolen fob can be remotely deactivated.

Two-Factor Authentication

Two-factor authentication requires two independent, simultaneous actions before access to a system is granted. The following are examples of two-factor authentication:

- In addition to a password, some systems require entering additional information known only to the authorized user, such as a mother's maiden name or the answer to another security question chosen by the authorized person. However, this security feature can be undermined if the secondary information can be obtained easily by an unauthorized third party.
- Passwords can be linked to biometrics.
- In addition to a password, a verification code is emailed or sent via text message that must be entered within a few minutes to complete the login.
- A biometric scan and a code from a fob are combined to allow access.

Considerations when evaluating the effectiveness of a logical data security system include:

- Does the system provide assurance that only authorized users have access to data?
- Is the level of access for each person appropriate to that person's needs?
- Is there a complete audit trail whenever access rights and data are modified?
- Are unauthorized access attempts denied and reported?

Other User Access Considerations

Besides user authentication, other security controls related to user access and authentication to prevent abuse or fraud include:

- **Automatic locking or logoff policies**. Any login that is inactive for a specified period of time can automatically be logged out in order to limit the window of time available for someone to take advantage of an unattended system.
- **Logs of all login attempts**, **whether successful or not**. Automatic logging of all login attempts can detect activities designed to gain access to an account by repeatedly guessing passwords. Accounts under attack can then be proactively locked in order to prevent unauthorized access.
- **Accounts that automatically expire**. If a user needs access to a system only for a short period of time, the user's access to the system should be set to automatically expire at the end of that period, thus preventing open-ended access.

Physical Access Controls

Physical access controls are used to secure equipment and premises. The goal of physical access controls is to reduce or eliminate the risk of harm to employees and of losing organizational assets. Controls should be identified, selected, and implemented based on a thorough risk analysis. Some common examples of general physical security controls include:

- Walls and fences
- Locked gates and doors
- Manned guard posts
- Monitored security cameras
- Guard dogs
- Alarm systems
- Smoke detectors and fire suppression systems

Physical access to servers and networking equipment should be limited to authorized persons. Keys are the least expensive way to manage access but also the weakest way because keys can be copied. A more effective method is **card access**, in which a magnetically encoded card is inserted into or placed near a reader. The card access also provides an audit trail that records the date, time, and identity of the person (or at least of the card of the person) who entered. One significant limitation of card access, however, is that a lost or stolen card can be used by anyone until it is deactivated.

Biometric access systems, discussed above as logical access controls, also serve as physical access controls. They can be used when physical security needs to be rigorous. Biometric access systems use physical characteristics such as blood vessel patterns on the retina, handprints, or voice authentication to authorize access. In general, such systems have a low error rate. That said, no single system is completely error-free, so biometric access systems are usually combined with other controls.

Controls can also be designed to **limit activities that can be performed remotely**. For example, changes to employee pay rates can be restricted to computers physically located in the payroll department. Thus, even if online thieves managed to steal a payroll password, they would be prevented from changing pay rates because they would not have access to the premises.

Technology-enabled Finance Transformation

Systems Development Life Cycle (SDLC)

The systems development life cycle was described previously in System Controls and Security Measures as *System and Program Development and Change Controls*, and it will be briefly reviewed here.

1) **Statement of objectives**. A written proposal is prepared, including the need for the new system, the nature and scope of the project and timing issues. A risk assessment is done to document security threats, potential vulnerabilities, and the feasible security and internal control safeguards needed to mitigate the identified risks.

2) **Investigation and feasibility study of alternative solutions**. Needs are identified and feasibility of alternative solutions are assessed, including the availability of required technology and resources. A cost-benefit analysis is done.

3) **Systems analysis**. The current system is analyzed to identify its strong and weak points, and the information needs from the new system, such as reports to be generated, database needs, and the characteristics of its operation are determined.

4) **Conceptual design**. Systems analysts work with users to create the design specifications and verify them against user requirements.

5) **Physical design**. The physical design involves determining the workflow, what and where programs and controls are needed, the needed hardware, backups, security measures, and data communications.

6) **Development and testing**. The design is implemented into source code, the technical and physical configurations are fine-tuned, and the system is integrated and tested. Data conversion procedures are developed.

7) **System implementation and conversion**. The site is prepared, equipment is acquired and installed, and conversion procedures, including data conversion, are implemented. System documentation is completed, procedures are developed and documented, and users are trained.

8) **Operations and maintenance**. The system is put into a production environment and used to conduct business. Continuous monitoring and evaluation take place to determine what is working and what needs improvement. Maintenance includes modifying the system as necessary to adapt to changing needs, replacing outdated hardware as necessary, upgrading software, and making needed security upgrades.

If follow-up studies indicate that new problems have developed or that previous problems have recurred, the organization begins a new systems study.

Business Process Analysis

Business process reengineering and redesign, discussed in Section D, *Cost Management*,[66] relies on technology to accomplish its objectives. When a business process needs to be reengineered or redesigned, new information systems will be needed.

For example, if the same input for a process is being keyed in more than once, such as into a database and also into a spreadsheet, that process needs to be redesigned so that the input is keyed in only one time. Not only is the duplication of effort wasteful, but inputting the information multiple times opens the process to the risk that the data will be keyed in differently each time. However, in order to successfully accomplish the redesign of the process, the information system being used will need to be redesigned, as well.

[66] See *Business Process Reengineering* and *Accounting Process Redesign*.

Any business operation, or process, consists of a group of related tasks or events that has a particular objective. **Business process analysis** is used to analyze a business process to determine the specific way the process is presently being accomplished from beginning to end. Business process analysis can provide information needed to monitor efficiency and productivity, locate process weaknesses, pinpoint potential improvements, and determine whether the potential improvements should be carried out.

Business process analysis involves the following steps:

1) **Determine the process to be analyzed**. Processes that influence revenue, expenses, the end product, and other critical functions are processes that should be analyzed periodically, as are processes that appear to be underperforming. A new process just implemented might also be analyzed to determine whether it is working as intended. Determine the beginning and ending points of the process to be analyzed.

2) **Collect information about the process that will be needed to analyze it**. Go through the documentation, interview the people involved, and do any other research necessary to answer any questions that arise.

3) **Map the process**. Business process mapping is visualizing the whole process from start to finish to better understand the various roles and responsibilities involved. Mapping makes it easier to see the big picture, what is working and what is not working, and where the risks are.

   A flowchart can be created for this step, or several software solutions, called business process analysis tools, are available for business process mapping.

4) **Analyze the process**. For example, determine the most important components and whether they could be improved. Look for any delays or other problems and determine whether they can be fixed. Look for ways to streamline the process so it uses fewer resources.

5) **Determine potential improvements**. Make recommendations for ways to improve the process. For example, determine whether incremental changes are needed and if so, what they are, or whether the process needs to be completely reengineered. Business process analysis tools can be an important part of this step, because they can be used to model changes to the process and prepare visuals.

Robotic Process Automation (RPA)

Robotic process automation (RPA), a type of artificial intelligence (see next topic), is **not** the same thing as the use of industrial robots. Robotic process automation software automates repetitive tasks by interacting with other IT applications to execute business processes that would otherwise require a human. RPA software can communicate with other systems to perform a vast number of repetitive jobs, and it can operate around the clock with no human errors. The automation of the repetitive part of a job frees up employees to do other things. RPA software cannot replace an employee, but it can increase the employee's productivity.

For example, RPA can be used when a change in policy necessitates a change in processing that would otherwise require additional employee time to implement or when sales growth causes changes in a system that would require either costly integration with another system or employee intervention.

The RPA software is not part of the organization's IT infrastructure. Instead, it interacts with the IT infrastructure, so no change is needed to the existing IT systems. Thus, RPA allows the organization to automate what would otherwise be a manual process without changing the existing systems.

Note: Robotics process automation **allows users to create their own robots** that can perform high-volume, repeatable tasks of low complexity faster and more accurately than humans can.

The software robots, also called "clients" or "agents," can log into applications, move files, copy and paste items, enter data, execute queries, do calculations, maintain records and transactions, upload scanned documents, verify information for automatic approvals or rejections, and perform many other tasks.

- RPA can be used to automate portions of transaction reporting and budgeting in the accounting area.
- RPA can automate manual consolidations of financial statements, leaving the accountants more time to follow up on items that require investigation, perhaps because of significant variances.
- Financial institutions can use RPA to automate account openings and closings.
- Insurance companies can use it to automate claims processing.
- RPA can be used in supply chain management for procurement, automating order processing and payments, monitoring inventory levels, and tracking shipments.

Any task that is high volume, rules-driven, and repeatable qualifies for robotic process automation. The RPA software can create reports and exception alerts so employees or management know when to get involved. The result can be cost savings, a higher accuracy rate, faster cycle times, and improved scalability if volumes increase or decrease. Employees can focus on more value-added activities.

Benefits of Robotic Process Automation

- Developing a process in RPA software does not require coding knowledge. RPA software usually has "drag-and-drop" functionality and simple configuration wizards that the user can employ to create an automated process. Some RPA solutions can be used to define a process by simply capturing a sequence of user actions.
- It enables employees to be more productive because they can focus on more advanced and engaging tasks, resulting in lower turnover and higher employee morale.
- It can be used to ensure that business operations and processes comply with regulations and standards.
- The tasks performed can be monitored and recorded, creating valuable data and an audit trail to further help with regulatory compliance as well as to support process improvement.
- Once an RPA process has been set up, the process can be completed much more rapidly.
- Robotic process automation can result in cost savings.
- It can help provide better customer service by automating customer service tasks. Customer service personnel can make use of it, or in some cases, RPA can even be used to converse with customers, gathering information and resolving their queries faster and more consistently than a person could.
- Low-volume or low-value processes that would not be economical to automate via other means can be automated using RPA.
- Business process outsourcing providers can use RPA tools to lower their cost of delivery or to offer "robots-as-a-service."
- Robots follow rules consistently, do not need to sleep, do not take vacations, do not get sick, and do not make typographical errors.

Limitations of Robotic Process Automation

- Robots are not infallible. Like any machine, their reliability is not 100%. Poor quality data input can cause exceptions, and their accuracy can be affected by system changes.
- Robots cannot replicate human reasoning. RPA software can mimic human behavior in the way it interacts with application user interfaces, but it can only follow highly methodical instructions.
- Robots have no "common sense." If a flaw in the instructions creates an error that would be obvious to a human, the robot will continue to follow the instructions provided without deviation and the error may be replicated hundreds or thousands of times before it is recognized by a human. Then, correcting all the incidents of the error could be very difficult (unless the errors could be corrected using the same automated tools).
- Because RPA can be used to automate processes in a "noninvasive" manner (in other words, without changing the IT system), management may be tempted to deploy RPA without relying on assistance from the IT department. However, although RPA **can** be deployed without involving the IT department, doing so may lead to unexpected problems. The IT department needs to be involved in the effort so the deployment is stable.

Artificial Intelligence (AI)

Artificial intelligence is a field in computer science dedicated to creating intelligent machines, especially computers, that can simulate human intelligence processes. AI uses **algorithms**, which are sets of step-by-step instructions that a computer can execute to perform a task. Some AI applications are able to learn from data and self-correct, according to the instructions given.

Artificial intelligence is categorized as either **weak AI**, also called **narrow AI,** or **strong AI**, also called **artificial general intelligence** (**AGI**).

- **Weak AI** is an AI system that can simulate human cognitive functions but although it appears to think, it is not actually conscious. A weak AI system is designed to perform a specific task, "trained" to act on the rules programmed into it, and it cannot go beyond those rules.
    - Apple's Siri[67] voice recognition software is an example of weak AI. It has access to the whole Internet as a database and is able to hold a conversation in a narrow, predefined manner; but if the conversation turns to things it is not programmed to respond to, it presents inaccurate results.
    - Industrial robots and robotic process automation are other examples of weak AI. Robots can perform complicated actions, but they can perform only in situations they have been programmed for. Outside of those situations, they have no way to determine what to do.
- **Strong AI** is equal to human intelligence and exists only in theory. A strong AI system would be able to reason, make judgments, learn, plan, solve problems, communicate, create and build its own knowledge base, and program itself. A strong AI system would be able to find a solution for an unfamiliar task without human intervention. It could theoretically handle all the same work that a human could, even the work of a highly-skilled knowledge worker.

Artificial intelligence is increasingly being used in administrative procedures and accounting. Robotic process automation, covered in the previous topic, is one application of AI. Other applications are **digital assistants** powered by AI and speech recognition (such as Siri), **machine vision**, and **machine learning**.

[67] Siri is a trademark of Apple Inc., registered in the U.S. and other countries.

Digital assistants have become standard in smartphones and for controlling home electronics, and their use has expanded into enterprise applications, as well. Some Enterprise Resource Planning systems incorporate digital assistants. The Oracle Cloud application includes functionality for enterprises to create **chatbots** and virtual assistants. A chatbot is software that can conduct a conversation via auditory or text methods. Examples of the use of chatbots are customer service and for acquisition of information.

Machine vision includes cameras, image sensors, and image processing software. It can automate industrial processes such as quality inspections by enabling robots to "see" their surroundings. Machine vision is also used in non-industrial settings such as surveillance and medical applications. It is increasingly being used in administrative and accounting applications, as well.

- Machine vision can be used to analyze satellite imagery for several purposes.
    - Insurance agents can use it to verify property information provided by existing clients or identify physical features of properties such as the roof condition and validate property features such as building size prior to providing an insurance quote to a new client, thereby reducing inspection costs.
    - Investment firms can use it to determine economic trends and forecast retail sales based on the number of cars in a retail parking lot on an hourly basis.
    - Financial institutions can monitor the status of construction on projects for construction lending purposes.
    - Businesses making investments in projects can use it to assess the degree to which a project is complete for accounting purposes and for monitoring and management.
- Machine vision can be used to automate document data extraction.
    - Businesses can assess large numbers of incoming paper documents or forms to extract the information from them. When documents are fed to the system, the software can identify each document as to its type and sort the documents to be forwarded to the appropriate processing group.
    - Incoming paper documents can be digitized for review by human employees, eliminating the need for manual data entry. Instead of doing the manual data input, the human employees can instead spend their time reviewing and ensuring the accuracy of fields entered by the machine learning software. The machine vision can even "read" handwritten text. When the data extraction is manually done, in many cases the full amount of the data is not input but only the most important data points are extracted. With machine vision, the data that results is complete and organized, and greater insights can be gained from data analytics.

Machine learning is another aspect of artificial intelligence being put to use in the accounting area. In machine learning, computers can learn by using algorithms to interpret data in order to predict outcomes and learn from successes and failures. Computers can "learn" to perform repeatable and time-consuming jobs such as the following.

- **Checking expense reports**. Computers can learn a company's expense reimbursement policies, read receipts, and audit expense reports to ensure compliance. The computer can recognize questionable expense reimbursement claims and forward them to a human to review.
- **Analyzing payments received on invoices**. When a customer makes a payment that needs to be applied to multiple invoices or that does not match any single invoice in the system, accounts receivable staff might need to spend time figuring out the proper combination of invoices to clear or may need to place a call to the customer, requiring considerable time and effort. However, a "smart machine" can analyze the possible invoices and match the payment to the right combination of invoices. Or, if the payment is short (is less than the amount due), the computer can apply the short payment and automatically generate an invoice for the remaining amount without any human intervention.

- **Risk assessment**. Machine learning can be used to compile data from completed projects to be used to assess risk in a proposed project.
- **Data analytics**. Using available data, machines can learn to perform one-time analytical projects such as how much the sales of a division have grown over a period of time or what the revenue from sales of a specific product was during a period of time.
- **Bank reconciliations**. Machines can learn to perform bank reconciliations.

AI-enabled robots will not replace accountants, but they will substantially transform what accountants do. When machines are able to do the repetitious work of calculating, reconciling, transaction coding, and responding to inquiries, accountants can focus less on tasks that can be automated and more on work such as advisory services that can be done only by humans, thereby increasing their worth in an organization. Accountants will need to monitor the interpretation of the data processed by AI to ensure that it continues to be useful for decision making. Accountants will need to embrace AI, keep their AI and analytical skills current, and be adaptive and innovative in order to remain competitive.

Considerations in Instituting Artificial Intelligence

- Processes should be re-imagined where possible, rather than just using the AI to replicate existing processes.
- Activities to be performed by AI should be those that are standardized and not often changed.
- Processes that are automated should be fully documented.
- Data quality, both input and output, must be reviewed. Potential exceptions and errors requiring human intervention must be identified and investigated.

Cloud Computing

Cloud computing is a method of essentially outsourcing the IT function. It is a way to increase IT capacity or add capabilities without having to invest in new infrastructure or license new software.

The National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce defines cloud computing as follows:

> Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.[68]

Thus, cloud computing means the use of business applications offered over the Internet. Cloud computing resources include data storage, infrastructure and platform (that is, hardware and operating system), and application software. Cloud service providers offer all three types of resources.

[68] *The NIST Definition of Cloud Computing*, Special Publication (NIST SP) Report Number 800-145, Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, September 2011, p. 2, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf, accessed April 22, 2019.

Software as a Service (**SaaS**) is defined by NIST as follows:

> The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.[69]

In other words, SaaS is software that has been developed by a cloud provider for use by multiple businesses (called **multi-tenant** use), and all business customers use the same software. Applications available as SaaS applications include enterprise resource planning (ERP), customer relationship management (CRM), accounting, tax and payroll processing and tax filing, human resource management, document management, service desk management, online word processing and spreadsheet applications, email, and many others.

Cloud computing also includes **Platform as a Service** (**PaaS**), and **Infrastructure as a Service** (**IaaS**).

NIST's definition of Platform as a Service is

> The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.[70]

If a company uses Platform as a Service, the company deploys its own applications to the cloud using the cloud provider's operating systems, programming languages, libraries, services, and tools. PaaS services include operating systems, database solutions, Web servers, and application development tools.[71]

Infrastructure in the context of cloud computing is both the hardware resources that support the cloud services being provided, including server, storage, and network components, and the software deployed.[72]

The definition of Infrastructure as a Service (IaaS) according to NIST is

> The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications and possibly limited control of select networking components (e.g., host firewalls).[73]

A company using Infrastructure as a Service is provided with physical and virtual processing, storage, networks, and other computing resources. The company can use the infrastructure to run software and operating systems. Although the company does not manage or control the cloud infrastructure, it does have control over the operating systems, storage, and deployed applications it uses, and it may have some control over things like configuration of a host firewall. Examples of Infrastructure as a Service include storage servers, network components, virtual machines, firewalls, and virtual local area networks.[74]

---

[69] Ibid.

[70] Ibid., pp. 2-3.

[71] *Moving to the Cloud*, Joseph Howell, *Strategic Finance* magazine, June 2015, © Institute of Management Accountants, https://sfmagazine.com/post-entry/june-2015-moving-to-the-cloud/, accessed April 22, 2019.

[72] *The NIST Definition of Cloud Computing*, Special Publication (NIST SP) Report Number 800-145, Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, September 2011, p. 2, note 2, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf, accessed April 22, 2019.

[73] Ibid., p. 3.

[74] *Moving to the Cloud*, Joseph Howell, *Strategic Finance* magazine, June 2015, © Institute of Management Accountants, https://sfmagazine.com/post-entry/june-2015-moving-to-the-cloud/, accessed April 22, 2019.

Benefits of Cloud Computing, SaaS, PaaS, and IaaS

- Users pay for only what they use, either on a periodic basis or on a usage basis. Thus, cloud computing is scalable. A firm can quickly increase or decrease the scale of its IT capability.
- Since the provider owns and operates the hardware and software, a user organization may be able to decrease its investment in its own hardware and software.
- The provider keeps the software updated, so the user organizations do not need to invest in upgrades or be concerned with applying them.
- Applications and data resident in the cloud can be accessed from anywhere, from any compatible device.
- Technology available in the cloud can be leveraged in responding to new and existing requirements for external compliance reporting, sustainability and integrated reporting, internal management reporting, strategic planning, budgeting and forecasting, performance measurement, risk management, advanced analytics, and many others.
- Cloud technology can be used to free up accountants so they can handle more higher-value activities and streamline lower-value processes.
- The cloud can enable the CFO to move into a more strategic role instead of spending time on transactional activities.
- The cloud can provide greater redundancy of systems than an on-site IT department may be able to offer, particular for small to medium-sized entities that may not be able to afford backup systems.
- The cloud can offer to companies of all sizes the advanced computing power needed for advanced analytics, something that otherwise only the largest companies would be able to afford due to cost. As a result, small to medium-sized businesses can be better positioned to compete with much larger competitors.
- Although security is a concern with the cloud, security is a concern with on-site IT, as well. The cloud frequently can provide stronger infrastructure and better protection than an on-site IT department may be able to.

Limitations, Costs, and Risks of Cloud Computing, SaaS, PaaS, and IaaS

- Reliability of the Internet is a concern. If the Internet goes down, operations stop.
- The quality of the service given by the provider needs to be monitored, and the service contract needs to be carefully structured.
- Loss of control over data and processing introduces security concerns. Selection of the cloud vendor must include due diligence. The cloud vendor must demonstrate that it has the proper internal controls and security infrastructure in place, and the vendor's financial viability as a going concern needs to be ascertained. Furthermore, the vendor's internal controls over data security and its infrastructure, as well as its continued viability as a going concern, need to be monitored on an ongoing basis.
- Contracting with overseas providers may lead to language barrier problems and time-zone problems as well as quality control difficulties.
- The ability to customize cloud solutions is limited, and that may hamper management from achieving all that it hopes to achieve.
- Service provided by automatic backup service providers may be problematic because timing of automatic backups may not be controllable by the user and may not be convenient for the user.

(Continued)

- The cloud cannot overcome weak internal controls. People are the greatest area of weakness with both internal IT and with cloud technologies. Security awareness training, proper hiring procedures, good governance, and protection from malware continue to be necessary after a company moves to the cloud, just as they are when the IT is on-site.
- The company's data governance must be structured to cover the cloud and the risks inherent in it, such as employees downloading new applications without authorization.
- Expected cost savings may not materialize. An organization may find that managing its own IT internally, even with all of its attendant problems, is less expensive than using the cloud.

Blockchains, Distributed Ledgers, and Smart Contracts

The concept of the blockchain and the first cryptocurrency, bitcoin, was first presented in 2009 in a paper titled *Bitcoin: A Peer-to-Peer Electronic Cash System*, ostensibly written by Satoshi Nakamoto. No one knows who Satoshi Nakamoto is, and it may be a pseudonym for one or more people.[75]

The blockchain was initially envisioned as a peer-to-peer system for sending online payments from one party to another party without using a financial institution. While online payments are still important, blockchain technology has expanded and is now used in many other areas.

Blockchain Terminology

Blockchain – A blockchain is a public record of transactions in chronological order, more specifically "a way for one Internet user to transfer a unique piece of digital property to another Internet user, such that the transfer is guaranteed to be safe and secure, everyone knows that the transfer has taken place, and nobody can challenge the legitimacy of the transfer."[76] Thus, a blockchain is a continuously growing digital record in the form of packages, called blocks, that are linked together and secured using cryptography. A blockchain is a system of digital interactions that does not need an intermediary such as a financial institution to act as a third party to transactions. Many users write entries into a record of information, the transactions are timestamped and broadcast, and the community of users controls how the record of information is updated. The digital interactions are secured by the network architecture of blockchain technology. The blocks are maintained via a peer-to-peer network of computers, and the same chain of blocks, called a ledger, is stored on many different computers.

Node – A node is a powerful computer running software that keeps the blockchain running by participating in the relay of information. Nodes communicate with each other to spread information around the network. A node sends information to a few nodes, which in turn relay the information to other nodes, and so forth.

Mining nodes, **or "miners"** – Miners are nodes (computers) on the blockchain that group outstanding transactions into blocks and add them to the blockchain.

Distributed ledger – A distributed ledger is a database held by each node in a network, and each node updates the database independently. Records are independently constructed and passed around the network by the various nodes—they are not held by any central authority. Every node on the network has information on every transaction and then comes to its own conclusion as to whether each transaction is authentic (that is, whether the people are who they say they are) and, if the transaction is a payment transaction, whether the sender has enough funds to cover the payment. The data sent in a transaction contains all the information needed to authenticate and authorize the transaction. When there is a consensus among the nodes that a transaction is authentic and should be authorized, the transaction is added to the distributed ledger, and all nodes maintain an identical copy of the ledger.

[75] For more information, see https://en.wikipedia.org/wiki/Satoshi_Nakamoto.

[76] Marc Andreessen, quoted on https://www.coindesk.com/information/what-is-blockchain-technology, accessed April 25, 2019.

Hash – Hashing is taking an input string of any length and giving it an output of a fixed length using a hashing algorithm. For example, bitcoin uses the hashing algorithm SHA-256 (Secure Hashing Algorithm 256) on Bitcoin networks. Any input, no matter how big or small, always has a fixed output of 64 symbols, which is made up of 256 bits, the source of the "256" in its name. The fixed output is the hash.

Block – A **block** is a record in a blockchain that contains and confirms many waiting transactions. It is a group of cryptocurrency transactions that have been encrypted and aggregated into the block by a miner. Each block has a header that contains (1) the details of the transactions in the block, including the senders' and receivers' addresses for each transaction and the amount of funds to be transferred from each sender to each receiver, (2) the hash of the information in the block just preceding it (which connects it to the blockchain), (3) a "**nonce**" (see below), and (4) the hash of the information in the block, including the nonce.

Nonce - The nonce in a block is a random string of characters that is appended to the transaction information in the block before the block is hashed and it is used to verify the block. After a nonce is added to the block, the information in the block, including the nonce, is hashed. The nonce needs to be a string of characters that causes the hash of the whole block, including the nonce, to conform to a particular requirement: the hash of the block that results after the nonce is included must contain a certain number of leading zeroes. If it does not, the nonce must be changed and the block hashed again.

The powerful mining nodes on the network all compete to determine the block's nonce by trying different values and re-hashing the block multiple times until one miner determines a nonce that results in the hash of the block conforming to the requirement for the number of leading zeroes. The mining node on the network that is the first to "solve the puzzle"—that is, calculate a nonce that results in a hash value for the block that meets the requirement for the number of leading zeroes—receives a reward of a certain number of units of the digital currency. (That is the source of the term "mining nodes." Receiving the currency reward is called "mining" the block because new digital currency is created and received.)

The blocks are hard to solve but easy to verify by the rest of the network once they are solved. Therefore, after one mining node solves the block, the other nodes on the network check the work to determine whether the block's nonce and its hash have been correctly calculated. If the other nodes agree that the calculations have been done correctly, the block is validated.

The determination of a nonce that fulfills the requirement for a new block's hash usually takes about 10 minutes and the nonce that fulfills the requirement is called "Proof of Work."

Note: "**Proof of Work**" is the **consensus algorithm** used on the Bitcoin blockchain. A proof of work is a piece of data that satisfies certain requirements, is costly and time-consuming to produce, but is easy for others to verify.

A consensus algorithm is a set of rules and number of steps for accomplishing a generally accepted decision among a group of people. For blockchains, consensus algorithms are used to ensure that a group of nodes on the network agree that all transactions are validated and authentic. The Proof of Work (PoW) algorithm is the most widely used consensus algorithm, but other blockchains may use different algorithms to accomplish the same thing.

Consensus algorithms are used to prevent double spending by a user, that is, spending the same digital currency more than once. In a traditional means of exchange, financial intermediaries ensure that currencies are removed from one account and placed into another, so that the same money is not spent more than once. However, digital currencies are decentralized and so there are no financial intermediaries. Consensus algorithms are used to make sure that all the transactions on the blockchain are authentic and that the currency moves from one entity to another entity. All the transactions are stored on the public ledger and can be seen by all participants, creating full transparency.

Confirmation – When a block has been validated on a blockchain, the transactions processed in it are **confirmed**. Confirmation means the transactions in the block have been processed by the network. Transactions receive a confirmation when they are included in a block and when each subsequent block is linked

to them. Once a transaction has received just one confirmation it is not likely to be changed because changing anything in the block would change the block's hash value. Since each block's hash value is part of the hash of the following block on the blockchain, changing anything in one block would require re-hashing all of the blocks following the changed block. Re-hashing all of the following blocks would necessitate recalculating a proper nonce for each subsequent block in turn.

If any change to a confirmed transaction is needed, a new transaction must be created.

Note: It is not impossible to change a previously-recorded transaction, but it would be very, very difficult to do so. Because of the time that would be required to recalculate all the nonces in all the subsequent blocks, it would be next to impossible to "catch up" to the most recent block, since new blocks would be being added to the chain all the time. Thus, transactions on a blockchain are considered **immutable**, which means "not subject or susceptible to change."

Immutability is important because if a transaction can be changed, a hacker could change the receivers' addresses and divert the payments, thus stealing a significant amount of the currency. That has actually happened to some of the smaller, less active, cryptocurrencies.

Bitcoin and Other Uses of Blockchain

The first usage of a blockchain was to transfer virtual currency, or cryptocurrency. A virtual currency is a digital representation of value that functions as a medium of exchange, a unit of account, and/or a store of value.[77] It is a piece of computer code that represents ownership of a digital asset.

Virtual currencies that have an equivalent value in real currency or that can act as a substitute for real currency are called "convertible virtual currency." Bitcoin was the first cryptocurrency and it is a convertible virtual currency. It can be digitally traded between users and can be purchased for and exchanged into U.S. dollars, Euros, and other currencies.[78]

The term "Bitcoin" is also used to refer to the protocol for the distributed ledger, that is, the distributed network that maintains a ledger of balances held in bitcoin tokens by users on the network. The word "bitcoin" with a lower-case "b" refers to the tokens, while the word "Bitcoin" with a capital "B" refers to the Bitcoin protocol and the Bitcoin network.

Note: Do not confuse the "distributed ledger" of a blockchain with an accounting ledger in which double-entry accounting is performed. When a blockchain is used to make payments, the transactions in the blockchain are single entry transactions for each entity and they represent the amount of cryptocurrency to be paid by the payor and received by the receiver. Although the transactions are not double-entry accounting entries for each entity, in a sense each currency transaction will be double entry since a sender's payment is equal to the receiver's receipt.

Other uses of blockchain include:

- Private, permissioned blockchains can be used by financial institutions for trading, payments, clearings, settlements, and repurchase agreement transactions (short-term borrowing of securities).[79]
- Intercompany transactions where different ERP systems are in use can be streamlined using a blockchain.
- Procurement and supply chain operations on blockchain can be used to optimize accounts payable or accounts receivable functions.

---

[77] *A CFTC Primer on Virtual Currencies,"* October 17, 2017, https://www.cftc.gov/LabCFTC/Primers/Index.htm, p. 4, accessed April 29, 2019.
[78] Ibid.
[79] *A CFTC Primer on Virtual Currencies,"* The U.S. Commodity Futures Trading Commission, October 17, 2017, p. 8, https://www.cftc.gov/LabCFTC/Primers/Index.htm, accessed April 29, 2019.

Smart Contracts

A contract that has been digitized and uploaded to a blockchain is called a **smart contract**. According to Nick Szabo, a computer scientist who envisioned smart contracts in 1996,

> "A smart contract is a set of promises, specified in digital form, including protocols within which the parties perform on these promises."[80]

A smart contract is created by translating the terms and conditions of a traditional agreement into a computational code written by blockchain developers in a programming language. It is basically a set of coded computer functions with a set of rules. The computer code is self-executing and performs an action at specified times or based on the occurrence or non-occurrence of an action or event such as delivery of an asset or a change in a reference rate. A simple example is a translation of "If X occurs, then Y makes a payment to Z."

A smart contract may include the elements of a binding contract (offer, acceptance, and consideration), or it may simply execute certain terms of a contract. When a smart contract is uploaded to a blockchain, the validity of the contract is checked and the required steps are enabled. After that, it is automatically executed.

A blockchain executes smart contracts in basically the same way as it executes transactions, and when the contract calls for payments to be made, they are made automatically in the cryptocurrency of the particular blockchain system being used. As with payments, the information in the contract is encrypted and hashed to a standardized size. A smart contract in a block of data is linked to the previous block. The smart contract is executed and payments are transferred according to its terms within the blockchain. The contract is funded by the payor so it can perform transactions automatically according to the agreement.

Following are some examples of the uses of smart contracts on blockchains.

- A blockchain can be used to ensure the authenticity of a product, so a purchaser of the product can be assured that the product he or she is buying is genuine and not a counterfeit. The information stored on the blockchain is unchangeable, so it is easier to prove the origins of a given product.
- It can be used to protect intellectual property. For example, artists can protect and sell music on a blockchain system. Artists who are due royalties from each sale of their material can receive the payments due them automatically through a smart contract as sales are made. They do not need to wait until the end of a period or wonder whether the publisher is being truthful about the number of sales made during the period.
- Blockchain and smart contracts have an important place in supply chain management, freight, and logistics, particularly in international transactions. Blockchain supply chain management does not rely on freight brokers, paper documents, or banks around the world to move goods and payments. The blockchain can provide secure digital versions of all documents that can be accessed by all the parties to a transaction. Defined events cause execution of the contract when the requirements are fulfilled. The smart contract can manage the flow of approvals and make the transfers of currency upon approval. For example, after goods have been received, the payment to the shipper takes place automatically. The transaction has complete transparency from beginning to end.
- On-demand manufacturing can be performed by machines that are automated and running on a blockchain network. A design would be sent to a locked machine along with an order for a specific number of units. The contract would be executed, the machine would be unlocked to produce the correct number of units, the goods would be produced, and the machine would be locked again.

[80] Nick Szabo, *Smart Contracts: Building Blocks for Digital Markets*, 1996, http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart_contracts_2.html, accessed April 29, 2019.

- An insurance contract can be in the form of a smart contract. For example, an orchard owner is concerned about a freeze that could destroy the year's fruit crop. An insurance company offers insurance against a freeze through a self-executing smart contract. The orchard owner and the insurance company agree to the contract terms and digitally sign a smart contract that is uploaded to a blockchain. The orchard owner's periodic premium payments are automated, and the blockchain checks a third-party source such as the National Weather Service daily for a possible freeze event. If a freeze event occurs, payment is sent automatically from the insurance company to the orchard owner.

  Note: The third-party source for a smart contract is called an **oracle**.

- Fish are being tracked from their sources to consumers in world markets and restaurants to prevent illegal fishing.
- The title to real property can be transferred using a blockchain. The whole history of the property owners and all the buy and sell transactions can be maintained in a blockchain dedicated to that piece of property. The blockchain can provide consensus regarding the current owner of the property and the historic record of property owners. The blockchain technology makes changing the historical records prohibitively difficult and costly. As a result, a title agency—a third party—is not needed to research the property records each time a piece of property is sold. The owner of the property can be identified using public key cryptography.
- Blockchains can be used to store data on archaeological artifacts held in museums. If an artifact or artifacts are stolen, the museum could release data using the hash codes to law enforcement so they could prevent the export or sale of artifacts matching the descriptions of the stolen items. Furthermore, if an artifact has a blockchain record, each time it crossed a border or was sold to a new collector the record would be updated to show it was legitimate. Looted antiquities would have no record, and faked records could not be manufactured.

Governance for Smart Contracts

Good governance is important for smart contracts, which require ongoing attention and may require action and possible revision. Governance standards and frameworks are needed and appear to be in the early stages of development.

- Governance standards may assign responsibility for smart contract design and operation and establish mechanisms for dispute resolution.
- Standards may incorporate terms or conditions that smart contracts need to have in order to be enforceable.
- Standards could create presumptions regarding the legal character of a smart contract, depending on its attributes and manner of use.
- Good governance standards may help address the risks that smart contracts present.[81]§

[81] Ibid., p. 31.

Benefits of Smart Contracts

- Smart contracts can authenticate counter-party identities, the ownership of assets, and claims of right by using digital signatures, which are private cryptographic keys held by each party.
- Smart contracts can access outside information or data to trigger actions, for example, commodity prices, weather data, interest rates, or an occurrence of an event.
- Smart contracts can self-execute. The smart contract will take an action such as transferring a payment without any action required by the counter-parties. The automatic execution can reduce counter-party risk and settlement risk.
- The decentralized, distributed ledger on the blockchain prevents modifications not authorized or agreed to by the parties.
- Smart contracts can enhance market activity and efficiency by facilitating trade execution.
- Use of standardized code and execution may reduce costs of negotiations.
- Automation reduces transaction times and manual processes.
- Smart contracts can perform prompt regulatory reporting.

Limitations and Risks of Smart Contracts

- The operation of a smart contract is only as smart as the information it receives and the computer code that directs it.
- Existing laws and regulations apply to all contracts equally regardless of what form a contract takes, so contracts or parts of contracts that are written in code are subject to otherwise applicable law and regulation. However, if a smart contract unlawfully circumvents rules and protections, to the extent it violates the law, it is not enforceable. For example, if a U.S. derivative contract is traded on or processed by a facility that is not appropriately registered with the U.S. Commodity Futures Trading Commission (CFTC) or is executed by entities required to be registered with the CFTC but which are not and do not have an exception or exemption from registration, the contract is prohibited.[82]
- A smart contract could introduce operational, technical, and cybersecurity risk.

  Operational risk: For example, smart contracts may not include sufficient backup and failover mechanisms in case of operational problems, or the other systems they depend on to fulfill contracts terms may have vulnerabilities that could prevent the smart contract from functioning as intended.

  Technical risk: For example, humans could make a typographical error when coding, Internet service can go down, user interfaces may become incompatible, the oracle (the third party source used by the smart contract to authorize payments) may fail or other disruptions can occur with the external sources used to obtain information on reference prices, events, or other data.

  Cybersecurity risk: Smart contract systems may be vulnerable to hacking, causing loss of digital assets. For example, an attacker may compromise the oracle, causing the smart contract to improperly transfer assets. There may be limited or no recourse if hackers transfer digital assets to themselves or others.[83]
- A smart contract may be subject to fraud and manipulation. For example, smart contracts can include deliberately damaging code that does not behave as promised or that may be manipulated. Oracles may be subject to manipulation or may themselves be fraudulent and may disperse fraudulent information that results in fraudulent outcomes.[84]

[82] *A Primer on Smart Contracts*, The U.S. Commodity Futures Trading Commission, Nov. 27, 2018, p. 25, https://www.cftc.gov/sites/default/files/2018-11/LabCFTC_PrimerSmartContracts112718_0.pdf, accessed April 29, 2019.
[83] Ibid, pp. 27-29.
[84] Ibid., p. 30.

Data Analytics

Data analytics is the process of **gathering and analyzing data in a way that produces meaningful information that can be used to aid in decision-making**. As businesses become more technologically sophisticated, their capacity to collect data increases. However, the stockpiling of data is meaningless without a method of efficiently collecting, aggregating, analyzing, and utilizing it for the benefit of the company.

Data analytics can be classified into four types: **descriptive analytics**, **diagnostic analytics**, **predictive analytics**, and **prescriptive analytics**.

1) **Descriptive analytics** report past performance. Descriptive analytics are the simplest type of data analytics and they answer the question, "What happened"?

2) **Diagnostic analytics** are used with descriptive analytics to answer the question, "Why did it happen"? The historical data is mined to understand the past performance and to look for the reasons behind success or failure. For example, sales data might be broken down into segments such as revenue by region or by product rather than revenue in total.

3) **Predictive analytics** focus on the future using correlative[85] analysis. Predictive analytics answer the question, "What is likely to happen"? Historical data is combined with other data using rules and algorithms. Large quantities of data are processed to identify patterns and relationships between and among known random variables or data sets in order to make predictions about what is likely to occur in the future. A sales forecast made using past sales trends is a form of predictive analytics.

4) **Prescriptive analytics** answer the question "What needs to happen?" by charting the best course of action based on an **objective** interpretation of the data. Prescriptive analytics make use of structured and unstructured data and apply rules to predict what will happen **and** to prescribe how to take advantage of the predicted events. For example, prescriptive analytics might generate a sales forecast and then use that information to determine what additional production lines and employees are needed to meet the sales forecast. In addition to anticipating what will happen and determining what needs to happen, prescriptive analytics can help determine **why** it will happen. Prescriptive analytics can incorporate new data and re-predict and re-prescribe, as well. Prescriptive analytics is most likely to yield the most impact for an organization, but it is also the most complex type of analytics.

Business Intelligence (BI)

Business intelligence is the combination of architectures, analytical and other tools, databases, applications, and methodologies that enable interactive access—sometimes in real time—to data such as sales revenue, costs, income, and product data. Business intelligence provides historical, current, and predicted values for internal, structured data regarding products and segments. Further, business intelligence gives managers and analysts the ability to conduct analysis to be used to make more informed strategic decisions and thus optimize performance.

The business intelligence process involves the transformation of data into information, then to knowledge, then to insight, then to strategic decisions, and finally to action.

- **Data** is facts and figures, but data by itself is not information.
- **Information** is data that has been processed, analyzed, interpreted, organized, and put into context such as in a report, in order to be meaningful and useful.

[85] If two things are correlated with one another, it means there is a close connection between them. It may be that one of the things causes or influences the other, or it may be that something entirely different is causing or influencing both of the things that are correlated.

- **Knowledge** is the theoretical or practical understanding of something. It is facts, information, and skills acquired through experience or study. Thus, information becomes knowledge through experience, study, or both.
- **Insight** is a deep and clear understanding of a complex situation. Insight can be gained through perception or intuition, but it can also be gained through use of business intelligence: data analytics, modeling, and other tools.
- The insights gained from the use of business intelligence lead to recommendations for the best action to take. **Strategic decisions** are made by choosing from among the recommendations.
- The strategic decisions made are implemented and turned into **action**.

A Business Intelligence system has four main components:

1) A data warehouse (DW) containing the source data.
2) Business analytics, that is, the collection of tools used to mine, manipulate, and analyze the data in the DW. Many Business Intelligence systems include artificial intelligence capabilities, as well as analytical capabilities.
3) A business performance management component (BPM) to monitor and analyze performance.
4) A user interface, usually in the form of a **dashboard**.

Note: A **dashboard** is a screen in a software application, a browser-based application, or a desktop application that displays in one place information relevant to a given objective or process, or for senior management, it may show patterns and trends in data across the organization.

For example, a dashboard for a manufacturing process might show productivity information for a period, variances from standards, and quality information such as the average number of failed inspections per hour. For senior management, it might present key performance indicators, balanced scorecard data, or sales performance data, to name just a few possible metrics that might be chosen.

A dashboard for a senior manager may show data on manufacturing processes, sales activity, and current financial metrics.

A dashboard may be linked to a database that allows the data presented to be constantly updated.

Big Data and the Four "V"s of Big Data

Big Data refers to vast datasets that are too large to be analyzed using standard software tools and so require new processing technologies. Those new processing technologies are **data analytics**.

Big Data can be broken down into three categories:

1) **Structured data** is in an organized format that enables it to be input into a relational database management system and analyzed. Examples include the data in CRM or ERP systems, such as transaction data, customer data, financial data, employee data, and vendor data.
2) **Unstructured data** has no defined format or structure. It is typically free-form and text-heavy, making in-depth analysis difficult. Examples include word processing documents, email, call center communications, contracts, audio and video, photos, data from radio-frequency identification (RFID) tags, and information contained on websites and social media.
3) **Semi-structured data** has some format or structure but does not follow a defined model. Examples include XML files, CSV files, and most server log files.

Big Data is characterized by four attributes, known as the **four V's**: volume, velocity, variety, and veracity.

1) **Volume**: Volume refers to the **amount** of data that exists. The volume of data available is increasing exponentially as people and processes become more connected, creating problems for accountants. The tools used to analyze data in the past—spreadsheet programs such as Excel and database software such as Access—are no longer adequate to handle the complex analyses that are needed. Data analytics is best suited to processing immense amounts of data.

2) **Velocity**: Velocity refers to the **speed** at which data is generated and changed, also called its **flow rate**. As more devices are connected to the Internet, the velocity of data grows and organizations can be overwhelmed with the speed at which the data arrives. The velocity of data can make it difficult to discern which data items are useful for a given decision. Data analytics is designed to handle the rapid influx of new data.

3) **Variety**: Variety refers to the diverse **forms** of data that organizations create and collect. In the past, data was created and collected primarily by processing transactions. The information was in the form of currency, dates, numbers, text, and so forth. It was **structured**, that is, it was easily stored in relational databases and flat files. However, today unstructured data such as media files, scanned documents, Web pages, texts, emails, and sensor data are being captured and collected. These forms of data are incompatible with traditional relational database management systems and traditional data analysis tools. Data analytics can capture and process diverse and complex forms of information.

4) **Veracity**: Veracity is the **accuracy** of data, or the extent to which it can be trusted for decision-making. Data must be objective and relevant to the decision at hand in order to have value for use in making decisions. However, various distributed processes—such as millions of people signing up online for services or free downloads—generate data, and the information they input is not subject to controls or quality checks. If biased, ambiguous, irrelevant, inconsistent, incomplete, or even deceptive data is used in analysis, poor decisions will result. Controls and governance over data to be used in decision-making are essential to ensure the data's accuracy. Poor-quality data leads to inaccurate analysis and results, commonly referred to as "garbage in, garbage out."

Some data experts have added two additional **V**s that characterize data:

5) **Variability**: Data flows can be inconsistent, for example, they can exhibit seasonal peaks. Furthermore, data can be interpreted in varying ways. Different questions require different interpretations.

6) **Value**: Value is the **benefit** that the organization receives from data. Without the necessary data analytics processes and tools, the information is more likely to overwhelm an organization than to help the organization. The organization must be able to determine the relative importance of different data to the decision-making process. Furthermore, an investment in Big Data and data analytics should provide benefits that are measurable.

Data Science

Data science is a field of study and analysis that uses algorithms and processes to extract hidden knowledge and insights from data. The **objective** of data science is to use both structured and unstructured data to extract information that can be used to develop knowledge and insights for forecasting and strategic decision making.

The difference between data analytics and data science is in their goals.

- The goal of data analytics is to provide information about issues that the analyst or manager either knows or knows he or she does not know (that is, "known unknowns").
- On the other hand, the goal of **data science** is to provide actionable insights into issues where the analyst or manager **does not know** what he or she does not know (that is, "unknown unknowns").

Example: Data science would be used to try to identify a future technology that does not exist today but that will impact the organization in the future.

Decision science, machine learning (that is, the use of algorithms that learn from the data in order to predict outcomes), and prescriptive analytics are three examples of means by which actionable insights can be discovered in a situation where "unknowns are unknown."

Data science involves data mining, analysis of Big Data, data extraction, and data retrieval. Data science draws on knowledge of data engineering, social engineering, data storage, natural language processing, and many other fields.

The size, value, and importance of Big Data has brought about the development of the profession of **data scientist**. Data science is a multi-disciplinary field that unifies several specialized areas, including statistics, data analysis, machine learning, math, programming, business, and information technology. A data scientist is a person with skills in all the areas, though most data scientists have deep skills in one area and less deep skills in the other areas.

Note: "Data mining" involves using algorithms in complex data sets to find patterns in the data that can be used to extract usable data from the data set. Data mining is discussed in more detail below.

Data and Data Science as Assets

Data and data science capabilities are strategic assets to an organization, but they are **complementary assets**.

- Data science is of little use without usable data.
- Good data cannot be useful in decision-making without good data science talent.

Good data and good data science, used together, can lead to large productivity gains for a company and the ability to do things it has never done before. Data and data science together can provide the following opportunities and benefits to an organization:

- They can enable the organization to make decisions based on data and evidence.
- The organization can leverage relevant information from various data sources in a timely manner.
- When the cloud is used, the organization can get the answers it needs using any device, any time.
- The organization can transform data into actionable insights.
- The organization can discover new opportunities.
- The organization can increase its competitive advantage.
- Management can explore data to get answers to questions.

The result can be maximized revenue, improved operations, and mitigated risks. The return on investment from the better decision-making that results from using data and data science can be significant.

As with any strategic asset, it is necessary to make investments in data and data science. The investments include building a modern business intelligence architecture using the right tools, investing in people with data science skills, and investing in the training needed to enable the staff to use the business intelligence and data analytics tools.

Challenges of Managing Data Analytics

Some general challenges of managing data analytics include data capture, data curation (that is, the organization and integration of disparate data collected from various sources), data storage, security and privacy protection, data search, data sharing, data transfer, data analysis, and data visualization.

In addition, some specific challenges of managing data analytics include:

- The growth of data and especially of unstructured data.
- The need to generate insights in a timely manner in order for the data to be useful.
- Recruiting and retaining Big Data talent. Demand has increased for data engineers, data scientists, and business intelligence analysts, causing higher salaries and creating difficulty filling positions.

Data Mining

Data mining is the use of statistical techniques to search large data sets to extract and analyze data in order to discover previously unknown, useful patterns, trends, and relationships within the data that go beyond simple analysis and that can be used to make decisions. Data mining uses specialized computational methods derived from the fields of statistics, machine learning, and artificial intelligence.

Data mining involves trying different hypotheses repeatedly and making inferences from the results that can be applied to new data. Data mining is thus an **iterative process**. **Iteration** is the repetition of a process in order to generate a sequence of outcomes. Each repetition of the process is a single iteration, and the outcome of each iteration is the starting point of the next iteration.[86]

Data mining is a process with defined steps, and thus it is a **science**. Science is the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence.[87]

Data mining is also an **art**. In data mining, decisions must be made regarding what data to use, what tools to use, and what algorithms to use. For example, one word can have many different meanings. In mining text, the **context** of words must be considered. Therefore, instead of just looking for words in relation to other words, the data scientist looks for whole phrases in relation to other phrases. The data scientist must make thoughtful choices in order to get usable results.

Data mining differs from statistics. Statistics focuses on explaining or quantifying the average effect of an input or inputs on an outcome, such as determining the average demand for a product based on some variable like price or advertising expenditures. Statistical analysis includes determining whether the relationships observed could be a result of the variable or could be a matter of chance instead. A simple example of statistical analysis is a linear regression model that relates total historical sales revenues (the dependent variable) to various levels of historical advertising expenditures (the independent variable) to discover whether the level of advertising expenditures affects total sales revenues. Statistics may involve using a sample from a dataset to make predictions about the population as a whole. Alternatively, it may involve using the entire dataset to estimate the best-fit model in order to maximize the information available about the hypothesized relationship in the population and predict future results.

In contrast, data mining involves open-ended exploring and searching within a large dataset without putting limits around the question being addressed. The goal is to predict outcomes for new individual records. The data is usually divided into a training set and a validation set. The training set is used to estimate the model, and the validation set is used to assess the model's predictive performance on new data.

Data mining might be used to classify potential customers into different groups to receive different marketing approaches based on some characteristic common to each group that is yet to be discovered. It may

[86] Definition of "iteration" from Wikipedia, https://en.wikipedia.org/wiki/Iteration, accessed May 8, 2019.

[87] Definition of "science" from the Science Council, https://sciencecouncil.org/about-science/our-definition-of-science/, accessed May 8, 2019.

be used to answer questions such as what specific online advertisement should be presented to a particular person browsing on the Internet based on their previous browsing habits and the fact that other people who browsed the same topics purchased a particular item.

Thus, data mining involves **generalization** of patterns from a data set. "Generalization" is the ability to predict or assign a label to a "new" observation based on a model built from past experience. In other words, the generalizations developed in data mining should be valid not just for the data set used in observing the pattern but should also be valid for new, unknown data.

Software used for data mining uses statistical models, but it also incorporates algorithms that can "learn" from the patterns in the data. An algorithm is applied to the historical data to create a mining model, and then the model is applied to new data to create predictions and make inferences about relationships. For example, data mining software can help find customers with common interests and determine which products customers with each particular interest typically purchase, in order to direct advertising messages about specific products to the customers who are most likely to purchase those products.

Data mining is used in **predictive analytics**. **Basic concepts** of predictive analytics include:

- **Classification** – Any data analysis involves classification, such as whether a customer will purchase or not purchase. Data mining is used when the classification of the data is not known. Similar data where the classification is known is used to develop rules, and then those rules are applied to the data with the unknown classification to predict what the classification is or will be. For example, customers are classified as predicted purchasers or predicted non-purchasers.
- **Prediction** – Prediction is similar to classification, but the goal is to predict the numerical value of a variable such as the **amount** of a purchase rather than (for example) simply classifying customers as predicted purchasers or predicted non-purchasers. Although classification also involves prediction, "prediction" in this context refers to prediction of a numerical value, which can be an integer (a whole number such as 1, 2, or 3) or a continuous variable.[88]
- **Association rules** – Also called **affinity analysis**, association rules are used to find patterns of association between items in large databases, such as associations among items purchased from a retail store, or "what goes with what." For example, when customers purchase a 3-ring notebook, do they usually also purchase a package of 3-hole punched paper? If so, the 3-hole punched paper can be placed on the store shelf next to the 3-ring notebooks.
- **Online recommendation systems** – In contrast to association rules, which generate rules that apply to an entire population, online recommendation systems use **collaborative filtering** to deliver personalized recommendations to users. Collaborative filtering generates rules for "what goes with what" at the individual user level. It makes recommendations to individuals based on their historical purchases, online browsing history, or other measurable behaviors that indicate their preferences, as well as other users' historical purchases, browsing, or other behaviors.
- **Data reduction** – Data reduction is the process of consolidating a large number of records into a smaller set by grouping the records into homogeneous groups.
- **Clustering** – Clustering is discovering groups in data sets that have similar characteristics without using known structures in the data. Clustering can be used in data reduction to reduce the number of groups to be included in the data mining algorithm.
- **Dimension reduction** – Dimension reduction entails reducing the number of variables in the data before using it for data mining, in order to improve its manageability, interpretability, and predictive ability.

[88] A continuous variable is a numerical variable that can take on any value at all. It does not need to be an integer such as 1, 2, 3, or 4, though it can be an integer. A continuous variable can be 8, 8.456, 10.62, 12.3179, or any other number, and the variable can have any amount of decimal points.

- **Data exploration** – Data exploration is used to understand the data and detect unusual values. The analyst explores the data by looking at each variable individually and looking at relationships between and among the variables in order to discover patterns in the data. Data exploration can include creating charts and dashboards, called data visualization or visual analytics (see next item). Data exploration can lead to the generation of a hypothesis.
- **Data visualization** – Data visualization is another type of data exploration. Visualization, or visual discovery, consists of creating graphics such as histograms and boxplots for numerical data in order to visualize the distribution of the variables and to detect outliers.[89] Pairs of numerical variables can be plotted on a scatter plot graph in order to discover possible relationships. When the variables are categorized, bar charts can be used. Visualization is covered in more detail later in this section.

Supervised and Unsupervised Learning in Data Mining

Supervised learning algorithms are used in classification and prediction. In order to "train" the algorithm, it is necessary to have a dataset in which the value of the outcome to be predicted is already known, such as whether or not the customer made a purchase. The dataset with the known outcome is called the **training data** because that dataset is used to "train" the algorithm. The data in the dataset is called **labeled** data because it contains the outcome value (called the **label**) for each record. The classification or prediction algorithm "learns" or is "trained" about the relationship between the predictor variables and the outcome variable in the training data. After the algorithm has "learned" from the training data, it is tested by applying it to another sample of labeled data for which the outcome is already known but is initially hidden (called the **validation data**) to see if it works properly. If several different algorithms are being tested, additional test data with known outcomes should be used with the selected algorithm to predict how well it will work. After the algorithm has been thoroughly tested, it can be used to classify or make predictions in data where the outcome is unknown.

Example: **Simple linear regression** is an example of a basic supervised learning algorithm. The *x* variable, the independent variable, serves as the predictor variable. The *y* variable, the dependent variable, is the outcome variable in the training and test data where the *y* value for each *x* value is known. The regression line is drawn so that it minimizes the sum of the squared deviations between the actual *y* values and the values predicted by the regression line. Then, the regression line is used to predict the *y* values that will result for new values of *x* for which the *y* values are unknown.[90]

Unsupervised learning algorithms are used when there is no outcome variable to predict or classify. Association rules, dimension reduction, and clustering are unsupervised learning methods.

Neural Networks in Data Mining

Neural networks are systems that can recognize patterns in data and use the patterns to make predictions using new data. Neural networks derive their knowledge from their own data by sifting through the data and recognizing patterns. Neural networks are used to learn about the relationships in the data and combine predictor information in such a way as to capture the complicated relationships among predictor variables and between the predictor variables and the outcome variable.

Neural networks are based on the human brain and mimic the way humans learn. In a human brain, neurons are interconnected and humans can learn from experience. Similarly, a neural network can learn from its mistakes by finding out the results of its predictions. In the same way as a human brain uses a network of neurons to respond to stimuli from sensory inputs, a neural network uses a network of artificial neurons,

[89] Outliers are data entries that do not fit into the model because they are extreme observations.
[90] Regression analysis is covered in more detail later in this section.

called **nodes**, to simulate the brain's approach to problem solving. A neural network solves learning problems by modeling the relationship between a set of input signals and an output signal.

The results of the neural network's predictions—the output of the model—becomes the input to the next iteration of the model. Thus, if a prediction made did not produce the expected results, the neural network uses that information in making future predictions.

Neural networks can look for trends in historical data and use it to make predictions. Some examples of uses of neural networks include the following.

- Picking stocks for investment by performing technical analysis of financial markets and individual investment holdings.
- Making bankruptcy predictions. A neural network can be given data on firms that have gone bankrupt and firms that have not gone bankrupt. The neural network will use that information to learn to recognize early warning signs of impending bankruptcy, and it can thus predict whether a particular firm will go bankrupt.
- Detecting fraud in credit card and other monetary transactions by recognizing that a given transaction is outside the ordinary pattern of behavior for that customer.
- Identifying a digital image as, for example, a cat or a dog.
- Self-driving vehicles use neural networks with cameras on the vehicle as the inputs.

The structure of neural networks enables them to capture complex relationships between predictors and an outcome by fitting the model to the data. It calculates weights for the individual input variables. The weights allow each of the inputs to contribute a greater or lesser amount to the output, which is the sum of the inputs. Depending on the effect of those weights on how well the output of the model—the prediction it makes—fits the actual output, the neural network then revises the weights for the next iteration.

Steps in Data Mining

A typical data mining project will include the following steps.

1) **Understand the purpose of the project**. The data scientist needs to understand the user's needs and what the user will do with the results. Also, the data scientist needs to know whether the project will be a one-time effort or ongoing.
2) **Select the dataset to be used**. The data scientist will take samples from a large database or databases, or from other sources. The samples should reflect the characteristics of the records of interest so the data mining results can be generalized to records outside of the sample. The data may be internal or external.
3) **Explore**, **clean**, **and preprocess the data**. Verify that the data is in usable condition, that is, whether the values are in a reasonable range and whether there are obvious outliers. Determine how missing data (that is, blank fields) should be handled. Visualize the data by reviewing the information in chart form. If using structured data, ensure consistency in the definitions of fields, units of measurement, time periods covered, and so forth. New variables may be created in this step, for example using the start and end dates to calculate the duration of a time period.
4) **Reduce the data dimension if needed**. Eliminate unneeded variables, transform variables as necessary, and create new variables. The data scientist should be sure to understand what each variable means and whether it makes sense to include it in the model.
5) **Determine the data mining task**. Determining the task includes classification, prediction, clustering, and other activities. Translate the general question or problem from Step 1 into the specific data mining question.
6) **Partition the data**. If supervised learning will be used (classification or prediction), partition the dataset randomly into three parts: one part for training, one for validation, and one for testing.
7) **Select the data mining techniques to use**. Techniques include regression, neural networks, hierarchical clustering, and so forth.

8) **Use algorithms to perform the task**. The use of algorithms is an iterative process. The data scientist tries multiple algorithms, often using multiple variants of the same algorithm by choosing different variables or settings. The data scientist uses feedback from an algorithm's performance on validation data to refine the settings.

9) **Interpret the results of the algorithm**. The data scientist chooses the best algorithm and tests the final choice on the test data to learn how well it will perform.

10) **Deploy the model**. The model is run on the actual records to produce actionable information that can be used in decisions. The chosen model is used to predict the outcome value for each new record, called **scoring**.[91]

A data mining project does not end when a particular solution is deployed, however. The results of the data mining may raise new questions that can then be used to develop a more focused model.

Challenges of Data Mining

Some of the challenges inherent in data mining include the following.

- **Poor data quality**. Data stored in relational databases may be incomplete, out of date, or inconsistent. For example, mailing lists can contain duplicate records, leading to duplicate mailings and excess costs. Poor quality data can lead to poor decisions.

  Furthermore, use of inaccurate data can cause problems for consumers. For example, when credit rating agencies have errors in their data, consumers can have difficulty obtaining credit.

- Information exists in **multiple locations** within the organization and thus is not centrally located, for example Excel spreadsheets that are in the possession of individuals in the organization. Information that is not accessible cannot be used.

- **Biases are amplified** in evaluating data. The meaning of a data analysis must be assessed by a human being, and human beings have biases. A "bias" is a preference or an inclination that gets in the way of impartial judgment. Most people tend to trust data that supports their pre-existing positions and tend not to trust data that does not support their pre-existing positions. Other biases include relying on the most recent data or trusting only data from a trusted source. All such biases contribute to the potential for errors in data analysis.

- Analyzed data often displays **correlations**.[92] However, **correlation does not prove causation**. Establishing a causal relationship is necessary before using correlated data in decision-making. If a causal relationship is assumed where none exists, decisions made on the basis of the data will be flawed.

- **Ethical issues** such as data privacy related to the aggregation of personal information on millions of people. Profiling according to ethnicity, age, education level, income, and other characteristics results from the collection of so much personal information.

- **Data security** is an issue because personal information on individuals is frequently stolen by hackers or even employees.

- A growing volume of **unstructured data**. Data items that are unstructured do not conform to relational database management systems, making capturing and analyzing unstructured data more complex. Unstructured data includes items such as social media posts, videos, emails, chat logs, and images, for example images of invoices or checks received.

[91] Shmueli, Galit, Bruce, Peter C., Yahav, Inbal, Patel, Nitin R., and Lichtendahl Jr., Kenneth C., *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*, 1st Edition, John Wiley & Sons, Hoboken, NJ, 2018, pp. 19-21.

[92] A "correlation" is a relationship between or among values in multiple sets of data where the values in one data set move in relation to the values in one or more other data set or sets.

Analytic Tools

Linear Regression Analysis

Regression analysis measures the extent to which an effect has historically been the result of a specific cause. If the relationship between the cause and the effect is sufficiently strong, regression analysis using historical data can be used to make decisions and predictions.

Time Series Analysis

Note: Time series analysis was introduced in Section B in Volume 1 of this textbook, topic B.3. *Forecasting Techniques*. Candidates may wish to review that information before proceeding. The trend pattern in time series analysis was introduced in *Forecasting Techniques* and will be further explained in this topic. Additional patterns will be discussed in this topic, as well.

A time series is a sequence of measurements taken at equally-spaced, ordered points in time. A time series looks at relationships between a variable and the passage of time. The variable may be sales revenue for a segment of the organization, production volume for a plant, expenses in one expense classification, or anything being monitored over time. Only one set of historical time series data is used in time series analysis and that set of historical data is not compared to any other set of data.

A time series can be descriptive or predictive. **Time series analysis** is used for **descriptive** modeling, in which a time series is modeled to determine its components, that is, whether it demonstrates a trend pattern, a seasonal pattern, a cyclical pattern, or an irregular pattern. The information gained from a time series analysis can be used for decision-making and policy determination.

Time series forecasting, on the other hand, is **predictive**. It involves using the information from a time series to forecast future values of that series.

A time series may have one or more of four patterns (also called **components**) that influence its behavior over time:

1) Trend
2) Cyclical
3) Seasonal
4) Irregular

Trend Pattern in Time Series Analysis

A trend pattern is the most frequent time series pattern and the one most amenable to use for predicting because the historical data exhibits a gradual shifting to a higher or lower level. If a long-term trend exists, short-term fluctuations may take place within that trend; however, the long-term trend will be apparent. For example, sales from year to year may fluctuate but overall, they may be trending upward, as is the case in the graph that follows.

A trend projection is performed with **simple linear regression analysis**, which forecasts values using historical information from all available past observations of the value. The regression line is called a **trend line** when the regression is being performed on a time series.

Note: The *x*-axis on the graph of a time series is always the horizontal axis and the *y*-axis is always the vertical axis. The *x*-axis represents the independent variable, also known as the predictor variable, and the *y*-axis represents the dependent variable, also known as the outcome variable.

In a time series regression analysis, the passage of time is the independent variable and is on the *x*-axis.

The equation of a simple linear regression line is:

$$\hat{y} = a + bx$$

Where:

$\hat{y}$ = the **predicted** value of $\hat{y}$ on the regression line corresponding to each value of *x, the* ***dependent variable****.*

a = the ***y*-intercept**, or the value of $\hat{y}$ on the regression line when *x* is zero, also called the **constant coefficient**.

b = the **slope** of the line and the amount by which the $\hat{y}$ value of the regression line changes (increases or decreases) when the value of *x* increases by one unit, also called the **variable coefficient**.

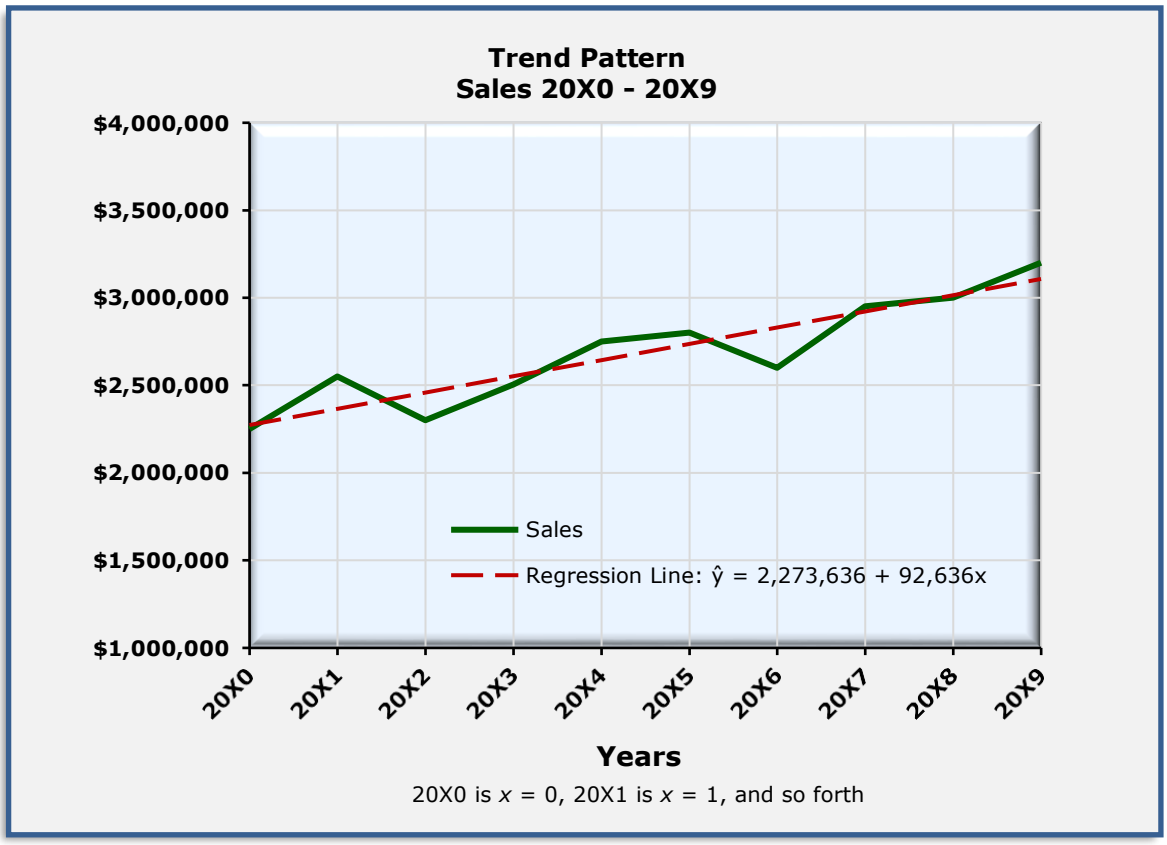
x = the **independent variable**, the value of *x* on the *x*-axis that corresponds to the predicted value of $\hat{y}$ on the regression line.

Example of a Trend Pattern in Time Series Analysis

Sales for each year, 20X0 through 20X9, are as follows:

Year	Sales
20X0	$2,250,000
20X1	$2,550,000
20X2	$2,300,000
20X3	$2,505,000
20X4	$2,750,000
20X5	$2,800,000
20X6	$2,600,000
20X7	$2,950,000
20X8	$3,000,000
20X9	$3,200,000

The following chart illustrates the trend pattern of the sales. It indicates a strong relationship between the passage of time (the *x* variable) and sales (the *y* variable) because the historical data points fall close to the regression line.



The regression equation, $\hat{y}$ = 2,273,636 + 92,636*x*, means the regression line begins at 2,273,636 and increases by 92,636 each succeeding year.

On the chart, 20X0 is at x = 0, 20X1 is at x = 1, and so forth.

The symbol over the "*y*" in the formula is called a "hat," and it is read as "*y*-hat." The *y*-hat indicates the **predicted value** of *y*, not the actual value of *y*. The predicted value of *y* **is the value of *y* on the regression line** (the line created from the historical data) at any given value of *x*.

Thus, in 20X4, where *x* = 4, the predicted value of *y*, that is, $\hat{y}$, is

$$\hat{y} = 2{,}273{,}636 + (92{,}636 \times 4)$$

$$\hat{y} = 2{,}273{,}636 + 370{,}544$$

$$\hat{y} = 2{,}644{,}180$$

In 20X7, where *x* = 7, the predicted value of *y* is

$$\hat{y} = 2{,}273{,}636 + (92{,}636 \times 7)$$

$$\hat{y} = 2{,}273{,}636 + 648{,}452$$

$$\hat{y} = 2{,}922{,}088$$

Those values for $\hat{y}$ (the value on the regression line) for 20X4 and 20X7 can be confirmed by looking at the chart.

Note: It would be a good idea to calculate the predicted value of y, that is, $\hat{y}$, at various other values of x and confirm the values on the graph. **The predicted value of *y* at a given value of *x* may be required in an exam question**.

The actual equation of the regression line as shown above **may not be given in an exam question**. The y-intercept and the slope of the regression line may be given instead. The constant coefficient, 2,273,636 in the above equation, is the y-intercept, and the variable coefficient, 92,636 in the above equation, is the slope of the line.

Thus, candidates need to know that the y-intercept of a regression line is the constant coefficient (the number that stands by itself) and the slope of the regression line is the variable coefficient (the number next to the x in the equation).

Furthermore, an exam question may use different letters to represent the variables and the coefficients, so candidates should be able to recognize the **form** of the equation and the meaning of each component of the formula based on its usage in the formula.

Trends in a time series analysis are not always upward and linear like the preceding graph. Time series data can exhibit an upward linear trend, a downward linear trend, a nonlinear (that is, curved) trend, or no trend at all. A scattering of points that have no relationship to one another would represent no trend at all.

Note: The CMA exam tests linear regression only.

Cyclical Pattern in Time Series Analysis

Any recurring fluctuation that lasts longer than one year is attributable to the **cyclical component** of the time series. A cyclical component in sales data is usually due to the cyclical nature of the economy.

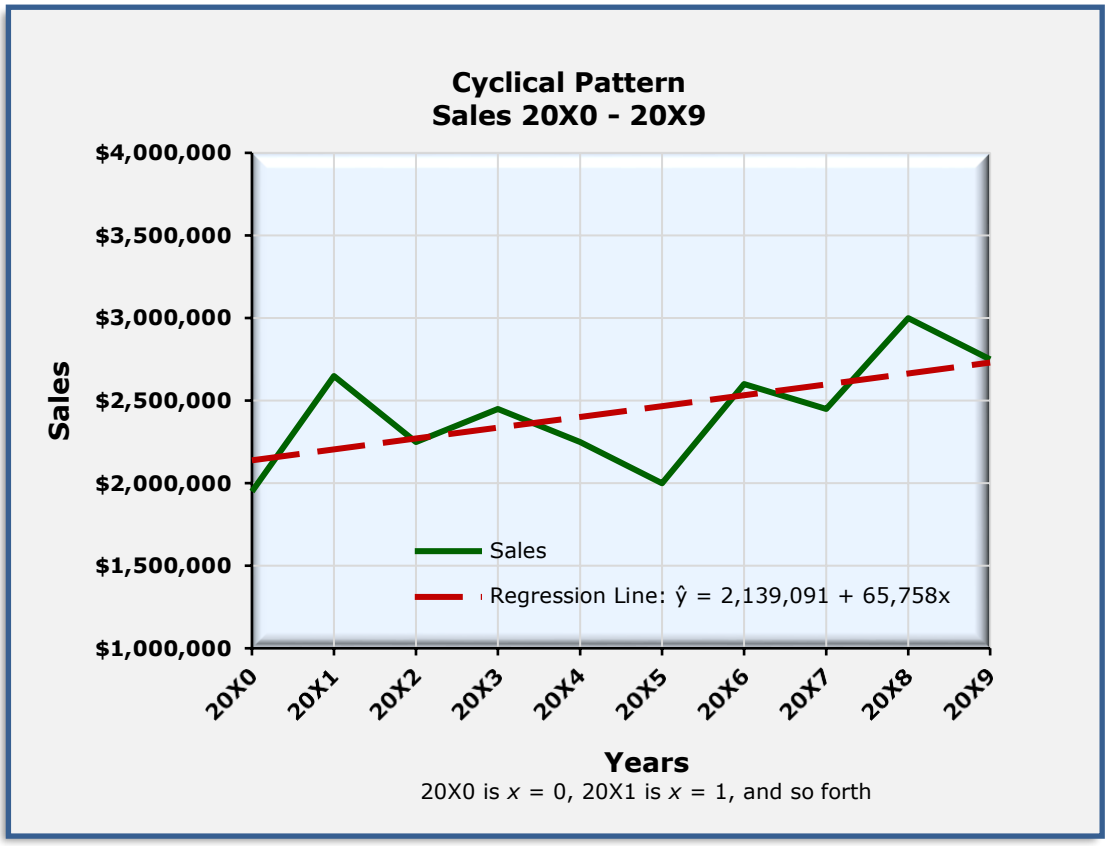
A long-term trend can be established even if the sequential data fluctuates greatly from year to year due to cyclical factors.

Example of a Cyclical Pattern in Time Series Analysis

Sales for each year, 20X0 through 20X9, are as follows:

Year	Sales
20X0	$1,975,000
20X1	$2,650,000
20X2	$2,250,000
20X3	$2,450,000
20X4	$2,250,000
20X5	$2,250,000
20X6	$2,600,000
20X7	$2,450,000
20X8	$3,000,000
20X9	$2,750,000

The following chart illustrates the cyclical pattern of the sales. The fluctuations from year to year are greater than they were for the chart containing the trend pattern. However, a long-term trend is still apparent.



Seasonal Pattern in Time Series Analysis

Usually, trend and cyclical components of a time series are tracked as annual historical movements over several years. However, a time series can fluctuate **within a year** due to seasonality in the business. For example, a surfboard manufacturer's sales would be highest during the warm summer months, whereas a manufacturer of snow skis would experience its peak sales in the wintertime. Variability in a time series due to seasonal influences is called the **seasonal component**.

Note: Seasonal behavior can take place within **any** time period. Seasonal behavior is not limited to periods of a year. A business that is busiest at the same time every day is said to have a **within-the-day seasonal component**. Any pattern that repeats regularly is a seasonal component.

Seasonality in a time series is identified by regularly spaced peaks and troughs with a consistent direction that are of approximately the same magnitude each time, relative to any trend. The graph that follows shows a strongly seasonal pattern. Sales are low during the first quarter each year. Sales begin to increase each year in the second quarter and they reach their peak in the third quarter, then they drop off and are low during the fourth quarter. However, the overall trend is upward, as illustrated by the trend line.

Example of a Seasonal Pattern in Time Series Analysis

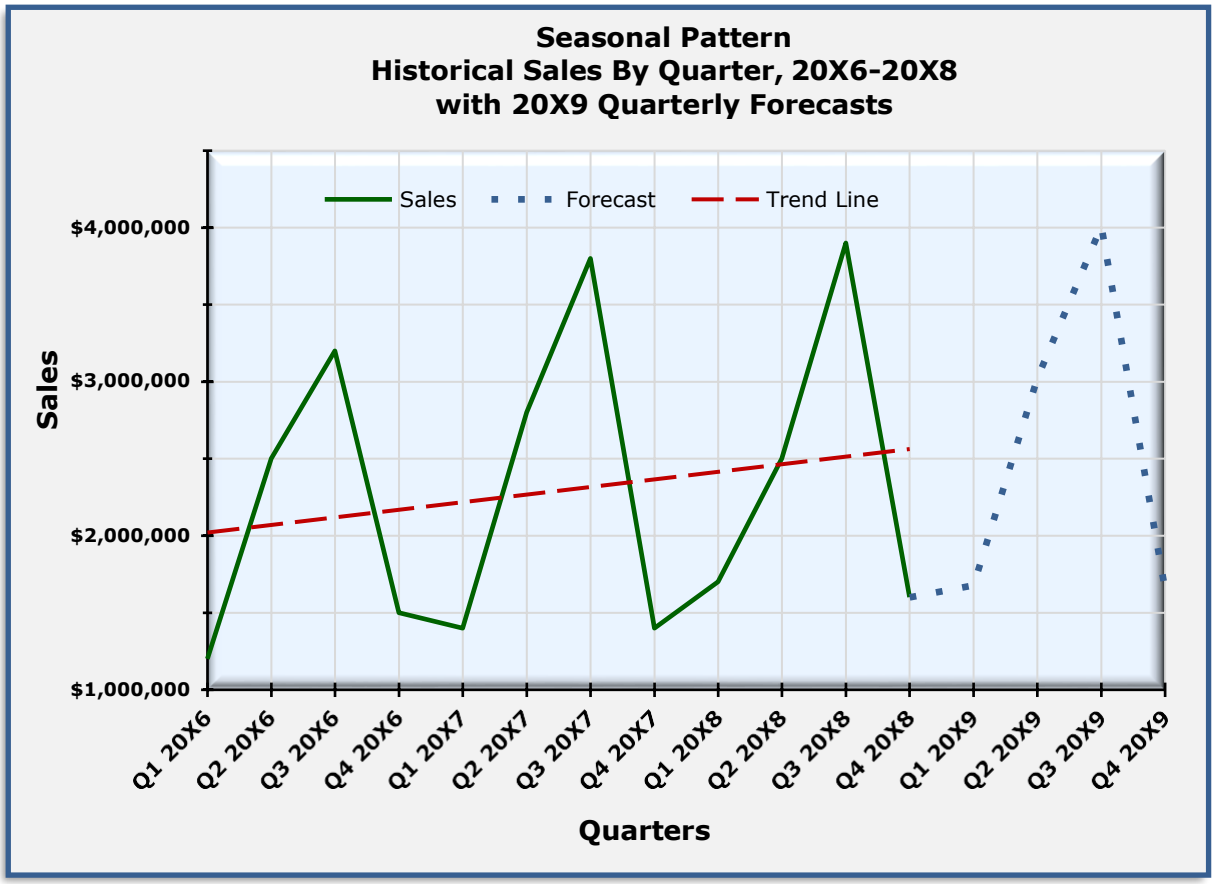
Sales for each quarter, March 20X6 through December 20X8, are as follows:

Year	Sales
Mar. 20X6	$1,200,000
Jun. 20X6	$2,500,000
Sep. 20X6	$3,200,000
Dec. 20X6	$1,500,000
Mar. 20X7	$1,400,000
Jun. 20X7	$2,800,000
Sep. 20X7	$3,800,000
Dec. 20X7	$1,400,000
Mar. 20X8	$1,700,000
Jun. 20X8	$2,500,000
Sep. 20X8	$3,900,000
Dec. 20X8	$1,600,000

The chart that follows contains historical sales by quarter for three years and forecasted sales by quarter for the fourth year. The fourth year quarterly forecasts were calculated in Excel using the FORECAST.ETS function, which is an exponential smoothing algorithm.

Note: Exponential smoothing is outside the scope of the CMA exams, so it is not covered any further in these study materials.

The chart illustrates that sales volume begins to build in the second quarter of each year. The sales volume reaches its peak in the third quarter and is at its lowest in the fourth quarter of each year.



Irregular Pattern in a Time Series

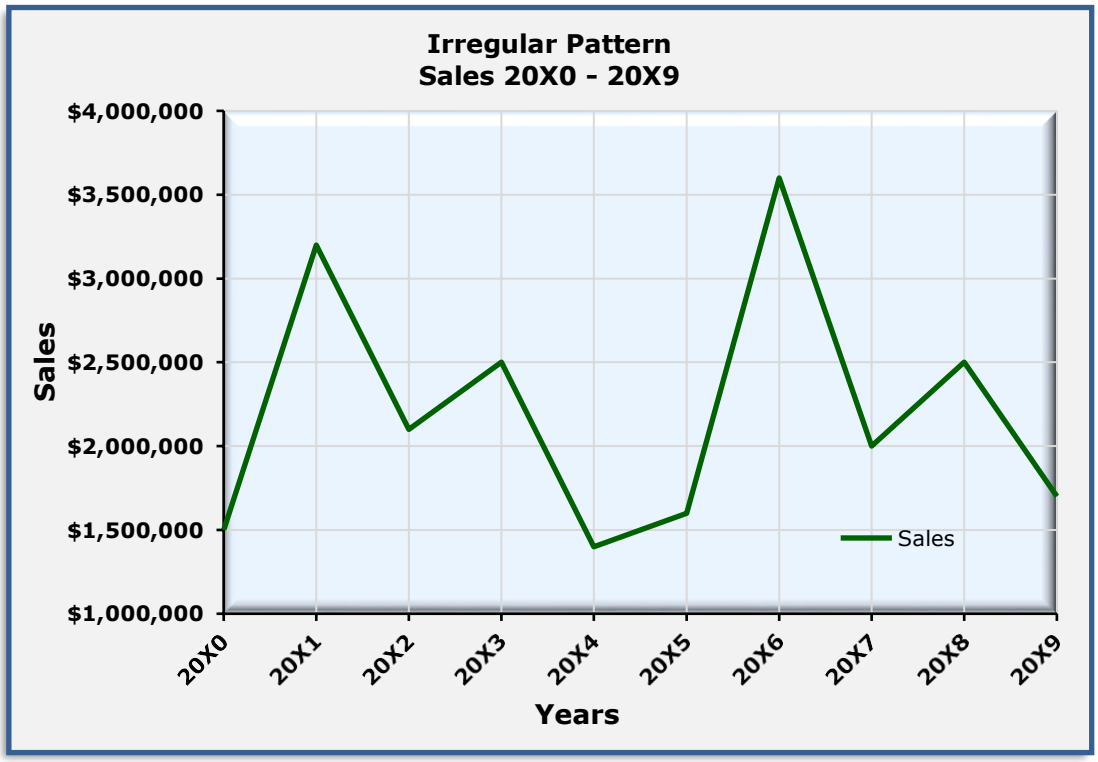
A time series may vary randomly, not repeating itself in any regular pattern. Such a pattern is called an **irregular pattern**. It is caused by short-term, non-recurring factors and its impact on the time series cannot be predicted.

Example of an Irregular Pattern in Time Series Analysis

Sales for each year, 20X0 through 20X9, are as follows:

Year	Sales
20X0	$1,500,000
20X1	$3,200,000
20X2	$2,100,000
20X3	$2,500,000
20X4	$1,400,000
20X5	$1,600,000
20X6	$3,600,000
20X7	$2,000,000
20X8	$2,500,000
20X9	$1,700,000

The following chart exhibits the irregular pattern of the sales:



Time Series and Regression Analysis

A time series that has a long-term upward or downward trend can be used to make a forecast. Simple linear regression analysis is used to create a trend projection and to forecast values using historical information from all available past observations of the value.

Note: Simple regression analysis is called "simple" to differentiate it from multiple regression analysis. The difference between simple linear regression and multiple linear regression is in the number of independent variables.

- A **simple** linear regression has only one independent variable. In a time series, that independent variable is the passage of time.
- A **multiple** linear regression has more than one independent variable.

Linear regression means the regression equation graphs as a straight line.

Simple linear regression analysis relies on two assumptions:

- Variations in the dependent variable (the value being predicted) are **explained by variations in one single independent variable** (the passage of time, for a time series).
- The relationship between the independent variable and the dependent variable (whatever is being predicted) is **linear**. A linear relationship is one in which the relationship between the independent variable and the dependent variable can be approximated by a straight line on a graph. The regression equation, which approximates the relationship, will graph as a straight line.

  The equation of a simple linear regression line is:

$$\hat{y} = a + bx$$

Where:

- $\hat{y}$ = the **predicted** value of the dependent variable, $\hat{y}$, on the regression line corresponding to each value of x.
- ***a*** = the **constant coefficient**, or the ***y*-intercept**, the value of $\hat{y}$ on the regression line when x is zero.
- ***b*** = the **variable coefficient** and the **slope of the regression line**, which is the amount by which the $\hat{y}$ value of the regression line changes (either increases or decreases) when the value of x increases by one unit.
- ***x*** = the **independent variable**, or the value of x on the x-axis that corresponds to the predicted value of $\hat{y}$ on the regression line.

Note: The equation of a simple linear regression line graphs as a straight line because none of the variables in the equation are squared or cubed or have any other exponents. If an equation contains any exponents, the graph of the equation will be a curved line.

The **line of best fit** as determined by simple linear regression is a formalization of the way one would fit a trend line through the graphed data just by looking at it. To fit a line by looking at it, one would use a ruler or some other straight edge and move it up and down, changing the angle, until it appears the differences between the points and the line drawn with the straight edge have been minimized. The line that results will be a straight line located at the position where approximately the same number of points are above the line as are below it and the distance between each point and the line has been minimized (that is, the distance is as small as possible).

Linear regression is used to calculate the location of the regression line mathematically. Linear regression analysis is performed on a computer or a financial calculator, using the observed values of x and y.

On a graph, the difference between each actual, observed point and its corresponding point on the calculated regression line is called a **deviation**. When the position of the regression line is calculated mathematically, the line will be in the position where **the deviations between each graphed value and the regression line have been minimized**. The resulting regression line is the **line of best fit**. That line can then be used to predict the value of y for any given value of x.

Note: The statistical method used to perform simple regression analysis is called the Least Squares, also known as the Ordinary Least Squares method or OLS. The regression line is called the **least squares regression line**.

Simple linear regression was used to calculate the regression line and the forecast on the graph presented earlier as an example of a **trend pattern**. The regression line was extended out for one additional year to create a forecast for that year.

Before Developing a Prediction Using Regression Analysis

Before using regression analysis to predict a value, determine whether regression analysis even can be used to make a prediction.

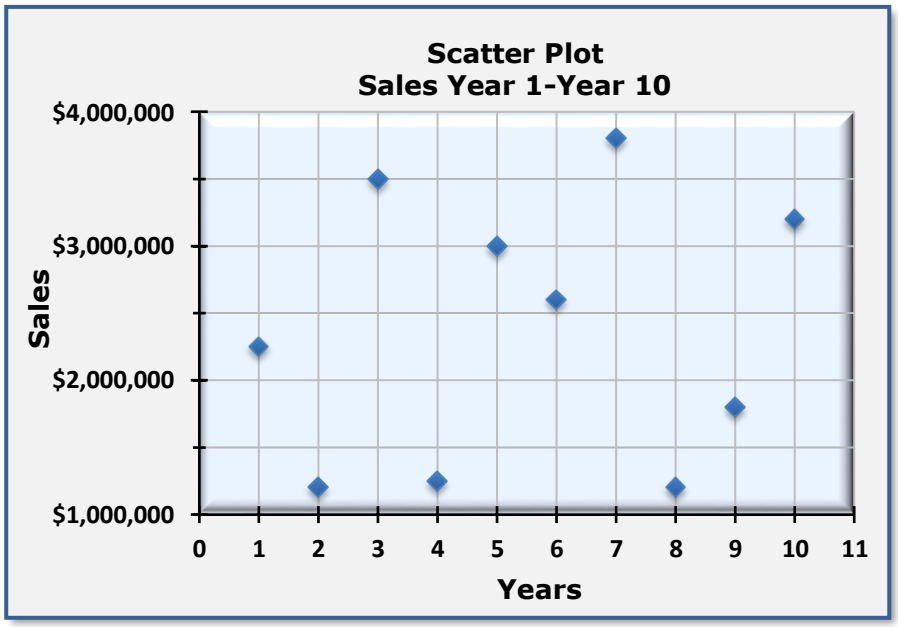
1) **The dependent variable, *y*, must have a linear relationship with the independent variable, *x*.**

   To determine whether a linear relationship exists, make a **scatter plot** of the actual historical values in the time series and review the results. Plotting the x and y coordinates on a scatter plot will indicate whether or not there is a linear relationship between them.

Note: The *x*-axis on a scatter plot is on the horizontal and the *y*-axis is on the vertical, and each observation is plotted at the intersection of its *x*-value and its *y*-value.

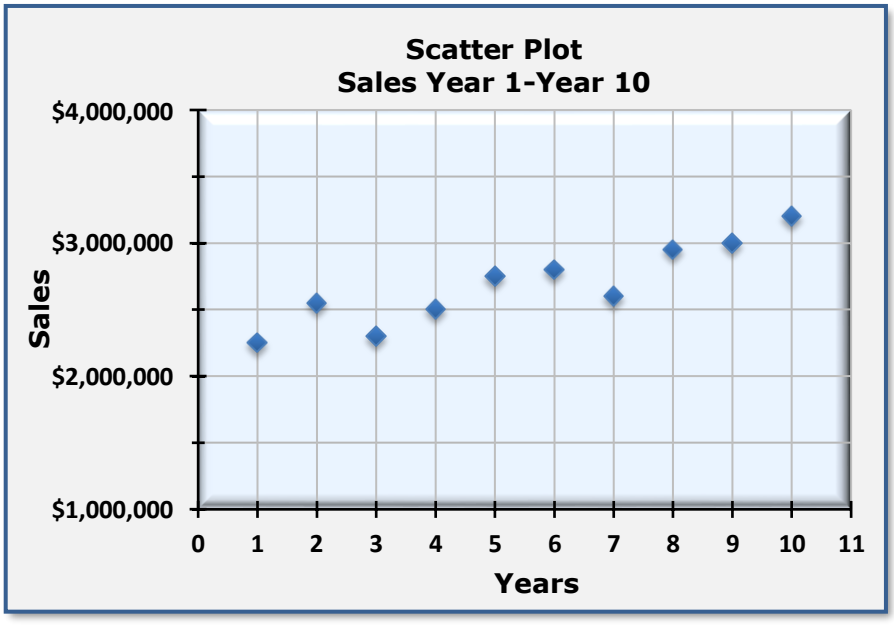
If the long-term trend appears to be linear, **simple linear regression analysis** may be able to be used (subject to correlation analysis as described below) to determine the location of the linear regression line, and that linear regression line can be used to make a prediction.

Below is a scatter plot that exhibits **no correlation** between the *x*-variable, time, and the *y*-variable, sales. The chart below is also an example of the **irregular pattern** described previously.



Historical sales like the above indicate that regression analysis using a time series would not be a good way to make a prediction.

On the other hand, the following scatter plot **does** display a linear relationship between the *x*-values and *y*-values, and the use of regression analysis to make a prediction could be helpful.



2) **Correlation analysis should indicate a high degree of correlation between the independent variable, *x*, and the dependent variable, *y*.**

In addition to plotting the points on a scatter plot graph, **correlation analysis** should be performed before relying on regression analysis to develop a prediction.

Correlation analysis determines the **strength of the linear relationship between the *x*-values and their related *y*-values**. The results of the correlation analysis tell the analyst whether the relationship between the independent variable (the passage of time, for a time series) and the dependent variable (sales, for example) is reasonable.

Note: Correlation describes the **degree of the relationship** between two variables. If two things are correlated with one another, it means there is a close connection between them.

- If high measurements of one variable tend to be associated with high measurements of the other variable, or low measurements of one variable tend to be associated with low measurements of the other variable, the two variables are said to be **positively correlated**.
- If high measurements of one variable tend to be associated with low measurements of the other variable, the two variables are said to be **negatively correlated**.
- If there is a close match in the movements of the two variables over a period of time, either positive or negative, it is said that the **degree of correlation is high**.

However, **correlation alone does not prove causation**. Rather than one variable causing the other variable to occur, it may be that some other, entirely different, factor is affecting both variables.

In a time series, the only independent variable is the passage of time. Many factors in addition to time can affect the dependent variable. For example, if sales are being predicted, economic cycles, promotional programs undertaken, and industry-wide conditions such as new government regulations can cause changes in sales volume. If time series regression analysis is used to develop a prediction, the prediction should be adjusted for other known factors that may have affected the historical data and that may affect the prediction.

Note: Correlation does not prove causation.

Correlation Analysis

Correlation analysis is used to assess how well a model can predict an outcome.

Correlation analysis involves several statistical calculations, all done with a computer or a financial calculator, using the observed values of *x* and *y*. Correlation analysis is used to determine how well correlated the variables are in order to decide whether the independent variable or variables can be used to make decisions regarding the dependent variable used in the analysis.

Some of the most important statistical calculations for determining correlation are:

1) The **correlation coefficient**, ***R***

2) The **standard error of the estimate**, also called the **standard error of the regression**

3) The **coefficient of determination, R^2**

4) The **T-statistic**

The Correlation Coefficient (*R*)

The **correlation coefficient** measures the relationship between the independent variable and the dependent variable. The coefficient of correlation is a number that expresses how closely related, or correlated,

the two variables are and the extent to which a variation in one variable has historically resulted in a variation in the other variable.

Mathematically, the correlation coefficient, represented by ***R***, is a numerical measure that expresses both the **direction** (positive or negative) and the **strength** of the linear association between the two variables. In a time series using linear regression analysis, the period of time serves as the independent variable (*x*-axis) while the variable such as the sales level serves as the dependent variable (*y*-axis).

When a time series (such as sales over a period of several years) is graphed, the data points on the graph may show an upsloping linear pattern, a downsloping linear pattern, a nonlinear pattern (such as a curve), or no pattern at all. The pattern of the data points indicates the amount of **correlation** between the values on the *x*-axis (time) and the values on the *y*-axis (sales).

The amount of correlation, or **correlation coefficient** (***R***), is expressed as a number **between −1 and +1**. The **sign** of the correlation coefficient and the **absolute value** of the correlation coefficient describe the **direction** and the **magnitude** of the relationship between the two variables.

- A correlation coefficient (*R)* of **+1** means the linear relationship between each value for *x* and its corresponding value for *y* is **perfectly positive** (upsloping). When *x* increases, *y* increases by the same proportion; when *x* decreases, *y* decreases by the same proportion.
- A correlation coefficient (*R*) of **−1** means the linear relationship between each value for *x* and its corresponding value for *y* is **perfectly negative** (downsloping). When *x* increases, *y* **decreases** by the same proportion; when *x* decreases, *y* **increases** by the same proportion. In a time series, the *x* value (time) can only increase, so if the correlation coefficient is negative, the level of the *y* value decreases with the passage of time.
- A correlation coefficient (*R*) that is **close to zero** usually means there is very little or no relationship between each value of *x* and its corresponding *y* value. However, a correlation coefficient that is close to zero may mean there is a strong relationship, but the relationship is not a linear one. (Candidates do not need to know how to recognize a non-linear relationship. Just be aware that non-linear relationships occur.)

A **high correlation coefficient** (*R*), that is, a number close to either +1 or −1, means that simple linear regression analysis would be useful as a way of making a projection. Generally, a correlation coefficient of ±0.50 or higher indicates enough correlation that a linear regression can be useful for forecasting. The closer *R* is to ±1, the better the forecast should be.

- If *R* is a positive number **close to +1** (such as 0.83), it indicates that the data points follow a linear pattern fairly closely and the pattern is upsloping. In a time series analysis of sales, it means sales are increasing as the time moves forward. A forecast made from this data using simple regression analysis should be fairly accurate.
- If *R* is a negative number **close to −1** (such as −0.77), it indicates in a time series that the data points follow a linear pattern, although less closely than in the previous example, and the pattern is downsloping instead of upsloping (for example, sales are decreasing as the time moves forward). A forecast made from this data using simple regression analysis would also be fairly accurate, though not as accurate as the previous example of an upsloping pattern, because the absolute number of 0.77 is lower than the absolute number of 0.83. In other words, −0.77 is further from −1 than 0.83 is from +1.

A **moderate correlation** coefficient (*R*), generally defined as ±0.30 to ±0.49, indicates a lower amount of correlation and questionable value of the historical data for forecasting.

A **low correlation coefficient** (*R*), around ±0.10, indicates that a forecast made from the data using simple regression analysis would not be useful.

The correlation coefficient (*R*) does not indicate **how much**, that is, the **proportion**, of the variation in the dependent variable that is explained by changes in the independent variable. The correlation coefficient

indicates only whether there is a direct (upsloping) or inverse (downsloping) relationship between the pairs of *x* and *y* variables and the strength of that relationship.

Note: The correlation coefficient, *R*, can be **calculated in Excel** by entering the *x* values in one column (for example, Column A, rows 1-10), the *y* values in another column (for example, Column B, Rows 1-10), and, for the current example, entering the following formula in any blank cell:

=CORREL(A1:A10,B1:B10)

The correlation coefficient is also a part of the output of a regression analysis performed in Excel.

Since Excel is not available on the CMA exam, candidates do not need to know how to calculate either a correlation coefficient or a regression analysis in Excel for the exam. The information will be given in any exam question where it is needed, and the candidate needs to know only how to interpret it.

Note: It is important to first **look at** the plotted data points on the scatter plot graph when determining whether or not there is a relationship between the independent variable and the dependent variable. Do not rely on the value of the correlation coefficient alone to indicate whether or not there is a relationship between the two variables because the correlation coefficient will not detect non-linear relationships.

Question 68: Correlation is a term frequently used in conjunction with regression analysis, and is measured by the value of the coefficient of correlation, R. The best explanation of the value R is that it

a) is always positive.

b) interprets variances in terms of the independent variable.

c) ranges in size from negative infinity to positive infinity.

d) is a measure of the relative relationship between two variables.

(CMA Adapted)

The Standard Error of the Estimate, also called the Standard Error of the Regression (*S*) and the Error Term (*e*)

Recall that the equation of a linear regression line, or the "line of best fit" on a graph is:

$$\hat{y} = a + bx$$

Where:

- $\hat{y}$ = the **predicted** value of the dependent variable, $\hat{y}$, on the regression line corresponding to each value of *x*.
- ***a*** = the **constant coefficient**, or the ***y*-intercept**, the value of $\hat{y}$ on the regression line when *x* is zero.
- ***b*** = the **variable coefficient** and the **slope of the regression line**, which is the amount by which the $\hat{y}$ value of the regression line changes (either increases or decreases) when the value of *x* increases by one unit.
- ***x*** = the **independent variable**, also called the **predictor variable**, or the value of *x* on the *x*-axis that corresponds to the value of $\hat{y}$ on the regression line.

Note: In a time series, the value of *x* only **increases** because time moves in only one direction: forward.

The equation of the simple linear regression model results in the **average**, or **predicted** value of $\hat{y}$ (the response) for any given value of *x* (the predictor). However, the actual observed data has responses that are not on the line itself, but rather they are **scattered around the regression line**. Thus, on a graph of a regression of historical data, there are two *y* values for each value of *x*: one is the actual historical value of *y* for that value of *x*, and the other is the estimated, or predicted, value of $\hat{y}$ for that value of *x*, represented by the *y* value on the trend line aligned with each value of *x*.

The scatter, that is, the difference between the actual value of *y* and the estimated value of $\hat{y}$ for each value of *x*, is called the **error term** or the **residual** for that value of *x*. The error term—the scatter of the data around the line—is represented by ***e*** in the linear regression model. If all of the historical data fell on the regression line, the error term for each value of *x* on the graph would be zero.

In algebra, an equation such as $y = 2{,}000 + 300x$ means that *y* is exactly $2{,}000 + 300x$. However, with regression data, the equation $\hat{y} = 2{,}000 + 300x$ is true **on average** but is not true for any given value of *x*.

Therefore, the equation that describes a **given actual**, **historical**, **value of *x*** used in a regression is as follows:

$$y = a + bx + e$$

Where:

y = the dependent variable (its actual, historical value, not its predicted value) corresponding to each value of *x.*

a = the **constant coefficient**, or the ***y*-intercept**, the value of $\hat{y}$ on the regression line when *x* is zero.

b = the **variable coefficient** and the **slope of the regression line**, or the average amount of change in *y* as a result of one unit of change in *x*.

x = the **independent variable**, or the value of *x* on the *x*-axis that corresponds to the value of $\hat{y}$ on the regression line.

e = the **error term**, also called the **residual**, which for each value of *x* is the difference between the estimated $\hat{y}$ value on the regression line for that value of *x* and the actual value of *y* for that value of *x.* **The error term will be different for each value of *x* used in the regression function**.

The **standard error of the estimate** (**S**) represents the **average distance that the observed values fall from the regression line**. In other words, it describes how **wrong** the regression model is **on average**, using the units of the dependent variable, *y*. The standard error of the estimate can give an indication of how well a linear regression works. It provides a comparison of the actual values of *y* (that is, the values that did occur historically) to the estimated values of $\hat{y}$ on the regression line. The estimated values of $\hat{y}$ on the regression line are the values that result from putting the various values for *x* into the regression function and calculating the resulting value of $\hat{y}$ at each value of *x*.

Each value of *x* has one residual, or error term. For any given value of *x*,

$$e = y - \hat{y}$$

Some of the residuals for a dataset that has been regressed will be positive and some will be negative. However, in a regression that has a constant term, **the mean of the residuals will be exactly zero**.

The standard error of the estimate (S) gives an indication of the precision, that is, the predictive ability of the regression model. The lower the standard error is, the more accurate will be the predictions made using the regression model.

Note: Smaller values for the standard error of the estimate are better because that indicates the observations are closer to the fitted line. However, the **size of the standard error of the estimate must be interpreted in relationship to the average size of the dependent variable**. If the average size of the dependent variable is 5,000,000 and the standard error is 250,000, the percentage of error is fairly small: 250,000 is only 5 percent of 5,000,000. If the size of the standard error of the estimate is less than 5 to 10 percent of the average size of the dependent variable, the regression analysis is fairly precise.

The inclusion of an error term in the regression model recognizes that:

- The regression model is imperfect.
- Some variables that help to "explain" the behavior of the dependent variable might not be included.
- The included variables may have been measured with error.

There is always some component in the variation of the dependent variable that is completely random.

The Coefficient of Determination (R^2)

The coefficient of determination is the **percentage of the total variation in the dependent variable** (y) **that can be explained by variations in the independent variable** (***x***)**, as depicted by the regression line**.

In a simple linear regression with only one independent variable, the **coefficient of determination is the square of the correlation coefficient**. The coefficient of determination is represented by the term R^2.

R^2 is expressed as a number between 0 and 1.

- If R^2 is 1, then 100% of the variation in the dependent variable is explained by variations in the independent variable.
- If R^2 is 0, then none of the variation in the dependent variable is explained by variations in the independent variable.

In a regression analysis with a high coefficient of determination (R^2), the data points will all lie close to the trend line. In a regression analysis with a low R^2, the data points will be scattered at some distances above and below the regression line. The higher the R^2, the better the predictive ability of the linear regression.

The T-Statistic

The t-statistic, or **t-value**, measures the degree to which the independent variable has a valid, long-term relationship with the dependent variable. The t-value for the independent variable used in a simple regression analysis **should generally be greater than 2**. A value below 2 indicates little or no relationship between the independent variable and the dependent variable, and thus the forecast resulting from the regression analysis should not be used.

Note: The equation of a linear regression line may be written in different ways, but *x* will usually represent the independent variable and *y* will usually represent the dependent variable.

The **constant coefficient** is the letter that stands by itself. The constant coefficient represents the *y*-intercept because it is the value of *y* when *x* is zero.

The **coefficient of the independent variable**, or the **variable coefficient**, is whatever term is next to the *x* in the formula. That term represents the **amount of change in *y* for each unit of increase in *x***, or the **slope** of the regression line.

The regression equation may be written in various ways, though the standard form of the equation used in statistics is

$$\hat{y} = a + bx$$

However, the equation $\hat{y} = ax + b$ is exactly the same as the equation $\hat{y} = a + bx$. The coefficients have just been expressed differently and the order of the terms on the right side of the equation is reversed. The right side of the equation may present the terms in any order.

If the regression line has a **negative** slope, that is, the values of $\hat{y}$ **decrease** for each increase in the value of *x*, the equation will be $\hat{y} = a - bx$.

Remember to **look for the *x***, or the independent variable. The term next to it will be the variable coefficient, or the amount of change in *y* for each unit of increase in *x*. The term that is all by itself will be the constant coefficient and the *y*-intercept, or the value of *y* when *x* is zero.

The symbol over the *y* in the formula is called a "hat," and it is read as "y-hat." The y-hat indicates the **predicted value** of *y*, not the actual value of *y*. The predicted value of *y* **is the value of *y* on the regression line** (the line created from the historical data) at any given value of *x*.

Multiple Regression Analysis

When more than one independent variable is known to impact a dependent variable and each independent variable can be expressed numerically, regression analysis using all of the independent variables to forecast the dependent variable is called **multiple** regression analysis.

Note: Remember that there must be a reasonable basis to assume a cause-and-effect relationship between the independent variable(s) and the dependent variable. If there is no reason for a connection, any correlation found through the use of regression analysis is accidental. **A linear relationship does not prove a cause-and-effect relationship, and correlation does not prove causation.**

The equation of a multiple regression function is usually written with either all "a"s or all "b"s as the coefficients, with a subscripted zero to indicate the constant coefficient and subscripted subsequent numerals to indicate the variable coefficients, such as the following, although any letters could be used:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \ldots + a_kx_k$$

Note: The variables and the coefficients in a multiple regression equation could be identified using **any** letters. To identify the various components, look for the **form** of the equation rather than the specific letters.

- The equation will have one component that stands by itself, and that will be the constant coefficient.
- The variable coefficients will be next to the independent variables.
- The independent variables may or may not be identified by "*x*"s.

Evaluating the Reliability of a Multiple Regression Analysis

In **simple** regression analysis, the **coefficient of determination**, R^2, is the proportion of the total variation in the dependent variable (y) that can be explained by variations in the independent variable (x). Thus, R^2 is an indicator of the reliability of a simple regression analysis.

R^2 is used as an indicator of the reliability of a **multiple** regression analysis, as well. In multiple regression analysis, though, the R^2 value evaluates the **whole** regression, including **all** of the independent variables used. If R^2 is above 0.50 or 50%, then the regression is fairly reliable because the regression can be used to predict that percentage of the total variation in the dependent variable. The higher the R^2 is, the better.

Multiple regression analysis uses the **t-value**—actually t-values (plural)—to evaluate the reliability of **each individual independent variable** as a predictor of the dependent variable. The t-value for each independent variable evaluates the **contribution** of that independent variable to the multiple regression analysis. A separate t-value is calculated for each of the individual independent variables in the multiple regression, and each independent variable is evaluated individually.

The t-value for every independent variable used in a multiple regression should generally be greater than 2, **the same as in simple regression analysis.** A t-value below 2 for a given independent variable indicates little or no relationship between that independent variable and the dependent variable; thus, that independent variable **should not be used in** the multiple regression.

As with simple regression, the R^2 and the t-values are based on the input supplied for the independent variables and the dependent variable.

Before using each independent variable in the multiple regression analysis, **each** independent variable being considered should be evaluated individually. Examine the output of the regression analysis for each independent variable, including the **coefficient of correlation**, ***R***, the **coefficient of determination**, $\boldsymbol{R^2}$, and the **t-value**. The evaluation will be similar to the evaluation done when with simple regression analysis with one independent variable.

If evaluation of each independent variable is done **before running the final multiple regression**, and if only those independent variables that are highly correlated with the dependent variable are used in the final multiple regression, the t-values that result from the final multiple regression should all be greater than 2. If, however, one or more of the t-values are below 2, then those independent variables should be eliminated and the multiple regression run again, even if the R^2 for the whole regression is high (for example, 0.85).

Note: Remember, correlation does not prove causation. In addition to a strong correlation between each independent variable and the dependent variable, there must be a logical cause-and-effect relationship between them before the independent variable can be used effectively to predict the dependent variable.

If the coefficient of determination, R^2, is only 0.50 for a multiple regression analysis (which is low but acceptable), the multiple regression might be used to predict 50% of the variation in the dependent variable, but only **if the t-values for all the independent variables are greater than 2**.

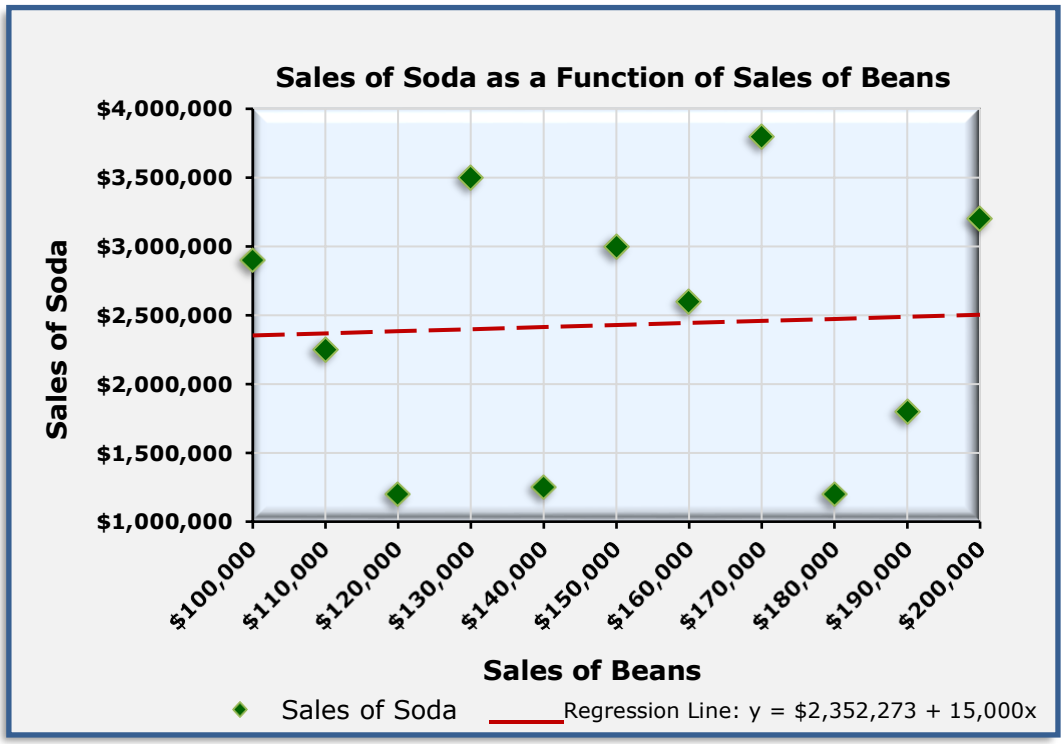
Goodness of Fit in Regression Analysis

The term **goodness of fit** describes how close the actual values used in a statistical model are to the expected values, that is, the predicted values, in the model.

In regression analysis, the regression equation is the model used to predict future values based on the behavior of the actual observations in response to a predictor. Thus, the correlation analysis described in this topic leads to the measurement of the regression equation's goodness of fit.

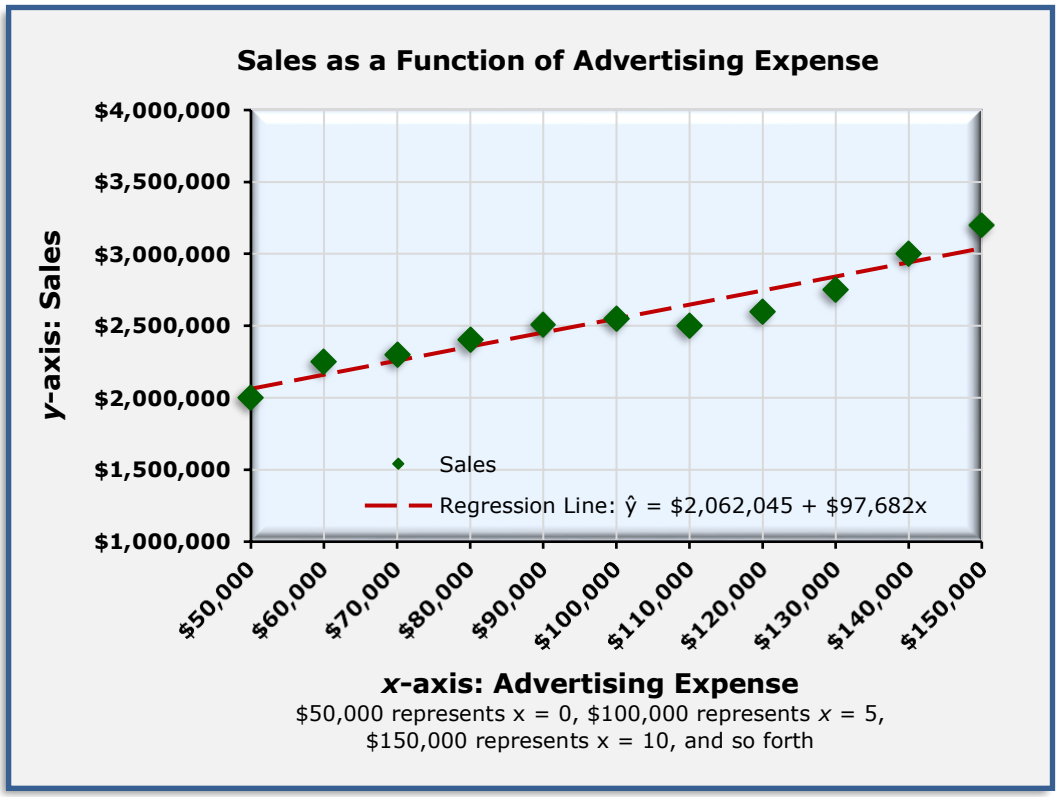
When the independent variable or variables used in the regression are **not** well correlated with the observations of the dependent variable used in the regression, the regression line is said to have a **low goodness of fit**.

The chart that follows exemplifies low goodness of fit for the regression equation, $\hat{y}$ = \$2,352,273 + \$15,000x. Sales of soda (the dependent variable on the y-axis) are regressed on sales of beans (the independent variable on the x-axis). Most of the points on the regression line are very far from the observed values of the dependent variable at that value for the independent variable, indicating very little correlation between the two variables.



On the other hand, when the independent variable or variables used in the regression are highly correlated with the observations of the dependent variable used in the regression, the regression equation is said to have a **high goodness of fit**.

Following is a chart showing sales revenue (on the *y*-axis) regressed on historical advertising expense (on the *x*-axis) for a representative period. The regression equation, $\hat{y}$ = $2,062,045 + $97,682*x*, has a high goodness of fit. The points representing historical sales as a function of advertising expense are very close to the regression line.



Remember, when calculating the predicted value of $\hat{y}$ at a given value of *x*, reassign values to the *x*-axis beginning with zero. For example, the value on the *x*-axis that represents advertising expense of $110,000 in the formula is *x* = 6. Thus, the predicted value of $\hat{y}$ when advertising expense is $110,000 is

$$\hat{y} = \$2,062,045 + (\$97,682 \times 6)$$

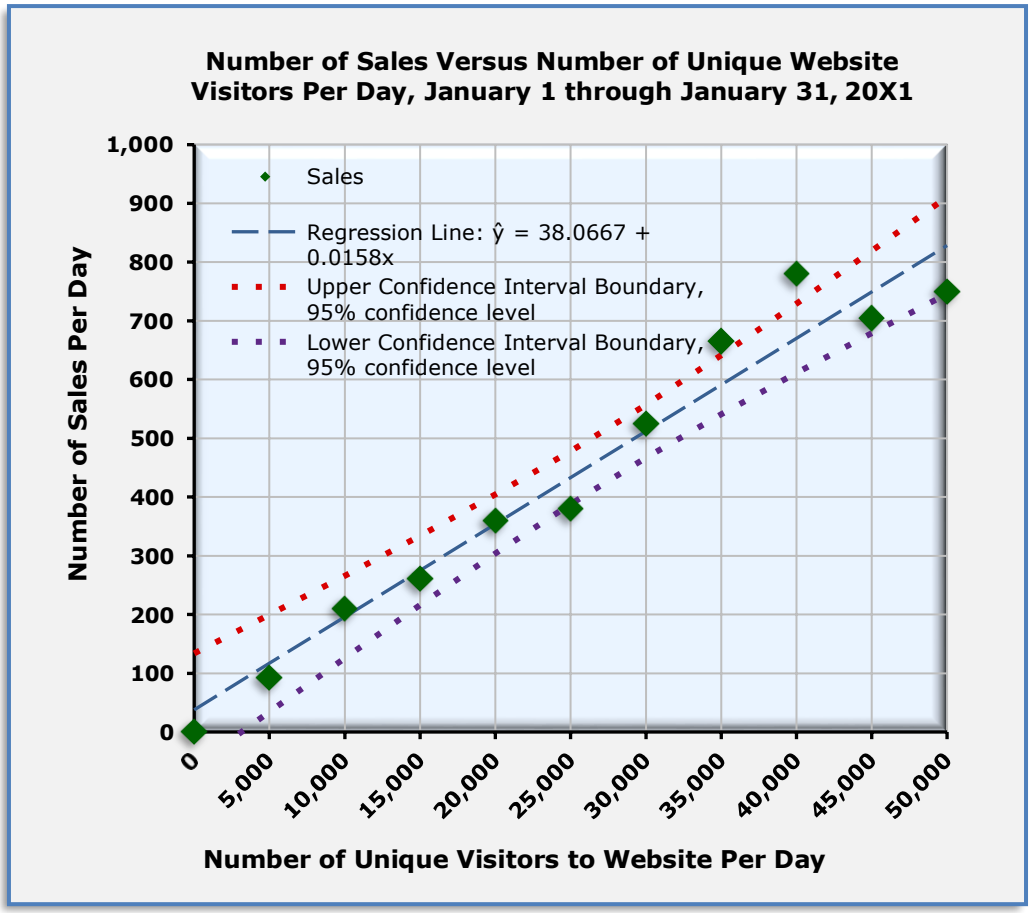
$$\hat{y} = \$2,648,137$$

In the above example, using $110,000 as *x* in the formula will result in an incorrect value for $\hat{y}$.

Confidence Interval in Regression Analysis

The **confidence interval** is used in regression analysis to describe the amount of uncertainty caused by the sampling method used when drawing conclusions about a population based on a sample. If several samples are drawn from a population using the same sampling method and a confidence interval at a confidence level of 95% is used, **95% of the interval estimates in the samples can be expected to include the true parameter of the population**.

The following chart contains information on number of sales made by an Internet retailer on its website regressed against the number of unique visitors to the website each day during a representative calendar period, January 1 through January 31, 20X1. The chart includes the upper and lower bands of a confidence interval at a 95% confidence level.



Note that several of the observations are actually outside the 95% confidence interval. That fact illustrates what the confidence interval is and highlights what it is not.

This sample's confidence interval at a confidence level of 95% does **not** mean that 95% of the observations in **this** sample, in any other sample, or in the population will be within the confidence interval, nor does it mean that the true value of sales as a function of website visitors will be within that interval 95% of the time. Instead, a confidence interval of 95% means that **if several periods are sampled and analyzed using the same 95% confidence interval**, **the proportion of those sample intervals that would contain the true number of sales to website visitors in the population would be equal to 95%**.

Example: If the same sampling of sales versus website visitors were performed for each month of 20X1, 20X2, and 20X3 (36 months) using the same 95% confidence interval, the location of the confidence interval bands would be slightly different for each month. However, **for 34 out of the 36 sampled months** (95%), the bands would contain the true value of sales related to website visitors for all past, present, and future periods (that is, the population).

Benefits and Limitations of Regression Analysis

Benefits of Regression Analysis
• Regression analysis is a quantitative method and as such it is objective. A given data set generates specific results. The results can be used to draw conclusions and make predictions. • Regression analysis is an important tool for drawing insights, making recommendations, and decision-making.
Limitations of Regression Analysis
• To use regression analysis, historical data are required. If historical data are not available, regression analysis cannot be used. • Even when historical data are available, the use of historical data is questionable for making predictions if a significant change has taken place in the conditions surrounding that data. • The usefulness of the data generated by regression analysis depends on the choice of independent variable(s). If the choice of independent variable(s) is inappropriate, the results can be misleading. • The statistical relationships that can be developed using regression analysis may be valid only for the range of data in the sample.

Sensitivity Analysis

Sensitivity analysis can be used to determine how much the prediction of a model will change if one input to the model is changed. It can be used to determine which input parameter is most important for achieving accurate predictions. Sensitivity analysis is known as "what-if" analysis.

To perform sensitivity analysis, define the model and run it using the base-case assumptions to determine the predicted output. Next, change one assumption at a time, leaving the other assumptions unchanged and run the model again to determine what effect changing that one assumption has on the predictions of the model. The amount of sensitivity of the prediction to the change in the input is the percentage of change in the output divided by the percentage of change in the input. Sensitivity analysis may reveal some area of risk that the company had not been aware of previously.

Monte Carlo Simulation Analysis

Whereas sensitivity analysis involves changing one input variable at a time, a **Monte Carlo** simulation analysis can be used to find solutions to mathematical problems that involve changes to multiple variables at the same time. Monte Carlo simulation can be used to develop an expected value when the situation is complex and the values cannot be expected to behave predictably. Monte Carlo simulation uses repeated random sampling and can develop probabilities of various scenarios coming to pass that can be used to compute a predicted result.

Adding a Monte Carlo simulation to a model allows the analyst to assess various scenario probabilities because various random values for the probabilistic inputs can be generated based on their probability distributions. The analyst determines ranges for the probabilistic inputs and also their probability distributions, means, and standard deviations. The application then generates the random values for the probabilistic inputs based on their ranges, probability distributions, means, and standard deviations as determined by the analyst.

The values for the probabilistic inputs are used to generate multiple possible scenarios, similar to performing statistical sampling experiments, except that it is done on a computer and over a much shorter time span than actual statistical sampling experiments. Enough trials are conducted (indeed, hundreds or thousands) with different values for the probabilistic inputs in order to determine a probability distribution for the resulting scenario, which is the output. The repetition is an essential part of the simulation.

For example, if the simulation is run to evaluate the probability that a new product will be profitable, the output may include a prediction for average profit and the probability of a loss.

Benefits of Sensitivity Analysis and Simulation Models

- Sensitivity analysis can identify the most critical variables, that is, the variables that are most likely to affect the end result if they are inaccurate. Since those are the variables that will make the most difference, those are the variables that should receive the most attention in making predictions.
- Simulation is flexible and can be used for a wide variety of problems.
- Both sensitivity analysis and simulation analysis can be used for "what-if" situations, because they enable the study of the interactive effect of variables.
- Both sensitivity analysis and simulation analysis are easily understood.
- Many simulation models can be implemented without special software packages because most spreadsheet packages provide useable add-ins. For more complex problems, simulation applications are available.

Limitations of Sensitivity Analysis and Simulation Models

- The results of sensitivity analysis can be ambiguous when the inputs used are themselves predictions.
- The variables used in a sensitivity analysis are likely to be interrelated. Changing just one variable at a time may fail to take into consideration the effect that variable's change will have on other variables.
- Simulation is not an optimization technique. It is a method that can predict how a system will operate when certain decisions are made for controllable inputs and when randomly generated values are used for the probabilistic inputs.
- Although simulation can be effective for designing a system that will provide good performance, there is no guarantee it will be the best performance.
- The results will be only as accurate as the model that is used. A poorly developed model or a model that does not reflect reality will provide poor results and may even be misleading.
- There is no way to test the accuracy of assumptions and relationships used in the model until a certain amount of time has passed.

Benefits of Data Analytics in General

- The process of cleaning the data preparatory to processing it can detect errors, duplicate information, and missing values. If the errors and duplicate information can be corrected and the missing values supplied, the data quality can be improved.
- The results of data analytics done correctly can lead to improved sales revenues and profits.
- It can help to reduce fraud losses by recognizing potentially fraudulent transactions and flagging them for investigation.
- Some easy-to-use data analytics tools are available that average users with little knowledge of data science are able to make use of to access data, perform queries, and generate reports. As a result, data scientists can be freed up to do more critical data analysis projects.
- Forecasting can be vastly improved through the use of data analytics.

Limitations of Data Analytics in General

- Big Data is used in data analytics to find correlations between variables. However, correlation does not prove causation. The fact that two variables are correlated does not mean that one variable caused the other. Both variables could have been caused by a third, unidentified, factor.
- Big Data can be used to find correlations and insights using an endless number of questions. But if the wrong questions are asked of the data, the answer will be meaningless even though it may be the "right" answer.
- Failure to take into consideration all relevant variables can lead to inaccurate predictions.
- Data breaches are a risk of using Big Data.
- Customer privacy issues and the risk of the misuse of data obtained from data analytics are matters for concern.
- In addition to the cost of the data analytics tools themselves, training on the use of the tools so they are used to their best advantage may entail costs, as well.
- Some easy-to-use data analytics tools are available that average users with little knowledge of data science are able to make use of to access data, perform queries, and generate reports. Use of the tools by those without a background in statistical analysis and data science and without adequate training, though, can cause risks such as data inconsistency, a lack of knowledgeable verification of the results, a lack of proper data governance, and ultimately, poor decisions.
- Selection of the right data analytics tools can be difficult.

Visualization

Data visualization is used for better understanding data and predictions from data. Charts, tables, and dashboards can be used to explore, examine, and display data. Interactive dashboards allow users to access and interact with real-time data and give managers a means to quickly see what might otherwise not be readily apparent. The choice of information to include in a dashboard depends on what a manager needs to see and can include visual presentations such as colored graphs showing, for instance, current customer orders.

In the data mining process, visualization is primarily used in exploration and cleaning of the data in the preprocessing step of data mining and in the reduction of the data dimensions step of the process.[93] For example:

- Visualization used in data exploration can help the analyst determine which variables to include in the analysis and which variables might be unnecessary.
- Visualization is used in data cleaning to find erroneous values in the historical data that need to be corrected (such as a sale recorded with a date 10 years in the future or a patient aged 250 years because his birth date is incorrect), missing values, duplicate records, columns that may have the same values in all the rows, and so forth.
- In data reduction, visualization can help in determining which categories can be combined.

Some visualization options are presented in the following pages, and information on how each can be used is provided.

Tables Used in Visualization

A table can be in any form and include all of the data available or only certain data.

The data table below will be used in all the chart examples that follow. The table contains data on the number of pounds of strawberries sold on each day of the week by a grocery store over a twelve-week period from June 3 through August 25. This information is used in placing orders, so that enough strawberries are purchased to meet the anticipated demand each day without over-buying and having excess strawberries that will need to be thrown away because they spoil.

In addition to daily strawberry sales for each of the days of the week for twelve weeks, the data table below contains the mean (average) for each day of the week over the twelve-week period.

Day of the Week	Jun. 3-9	Jun. 10-16	Jun. 17-23	Jun. 24-30	Jul. 1-7	Jul. 8-14	Jul. 15-21	Jul. 22-28	Jul.29-Aug. 4	Aug. 5-11	Aug. 12-18	Aug. 19-25	Mean
Mon.	15	10	28	39	48	25	12	20	30	23	28	22	25
Tues.	35	25	45	40	46	49	30	60	38	32	22	58	40
Wed.	68	42	57	74	84	55	30	55	60	75	88	32	60
Thur.	60	80	65	90	65	85	50	70	110	75	45	105	75
Fri.	95	60	85	90	70	80	105	85	50	80	105	55	80
Sat.	110	75	85	98	75	102	85	50	120	100	65	115	90
Sun.	11	7	14	10	40	18	20	25	35	20	22	18	20

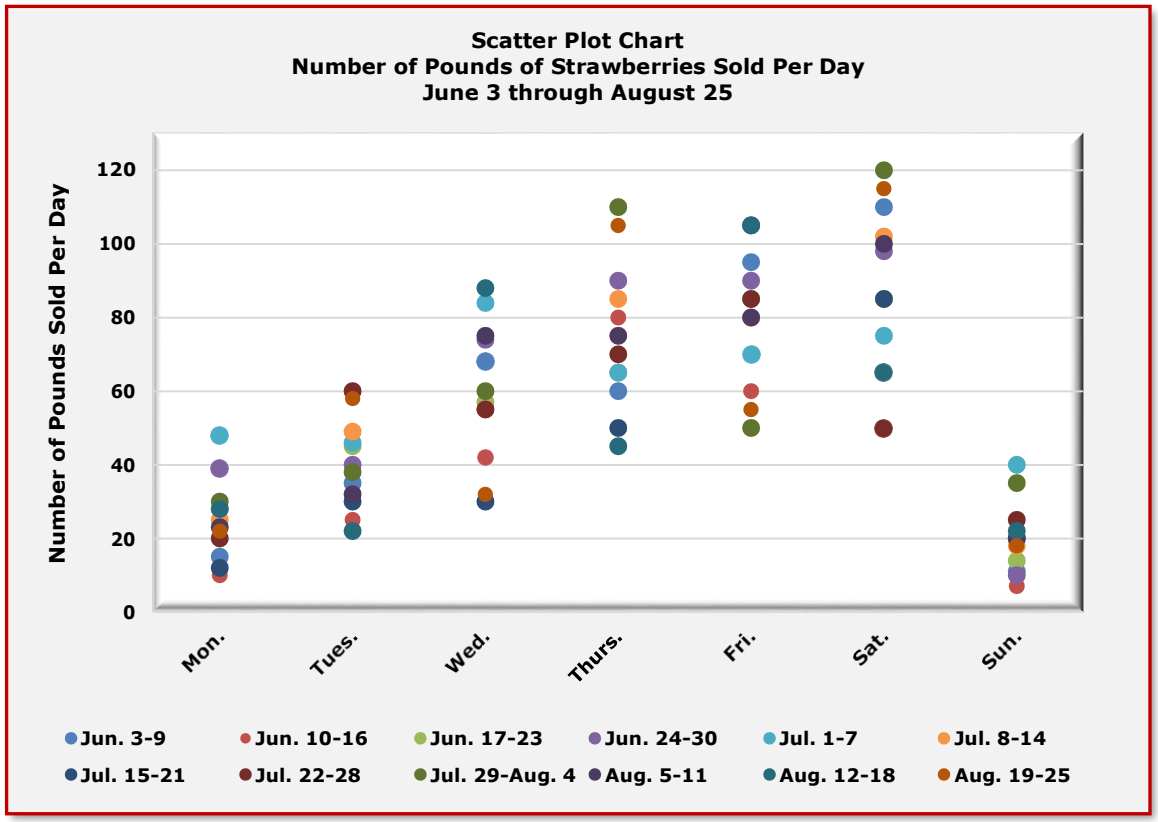
[93] See *Steps in Data Mining* earlier in this section.

Scatter Plot

A scatter plot can be used to show all the values for a dataset, typically when there are two variables. One variable may be independent and the other value dependent, or both variables may be independent. The independent variable is generally plotted on the horizontal (*x*) axis and the dependent variable is plotted on the vertical (*y*) axis.

A scatter plot can reveal correlations between variables or alternatively, a lack of correlation. For example, do sales of strawberries correlate with days of the week? A scatter plot can answer that question.

In this case, it appears that strawberry sales do correlate with days of the week. Sales build from Monday through Saturday and then they drop off on Sunday each week.

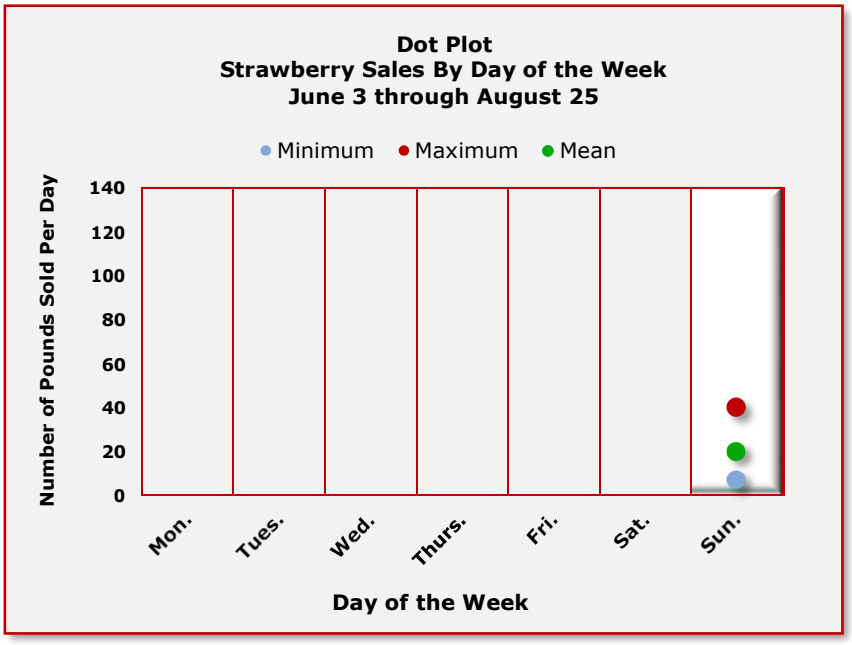


Charts Containing Summarized Statistics

Several charts are used to present summarized statistics such as means, maximum values, and minimum values.

Dot Plot

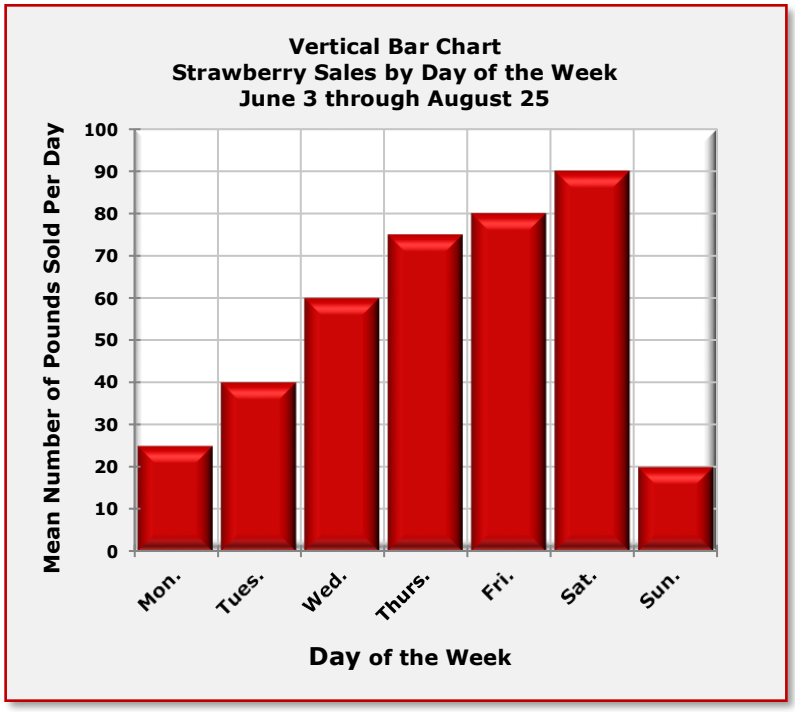
A dot plot provides information in the form of dots. A dot plot can be used to visualize several data points for each category on the *x*-axis. For example, the following dot plot shows the minimum, the maximum, and the mean number of pounds of strawberries sold for each day of the week during the twelve-week period.



Bar Chart

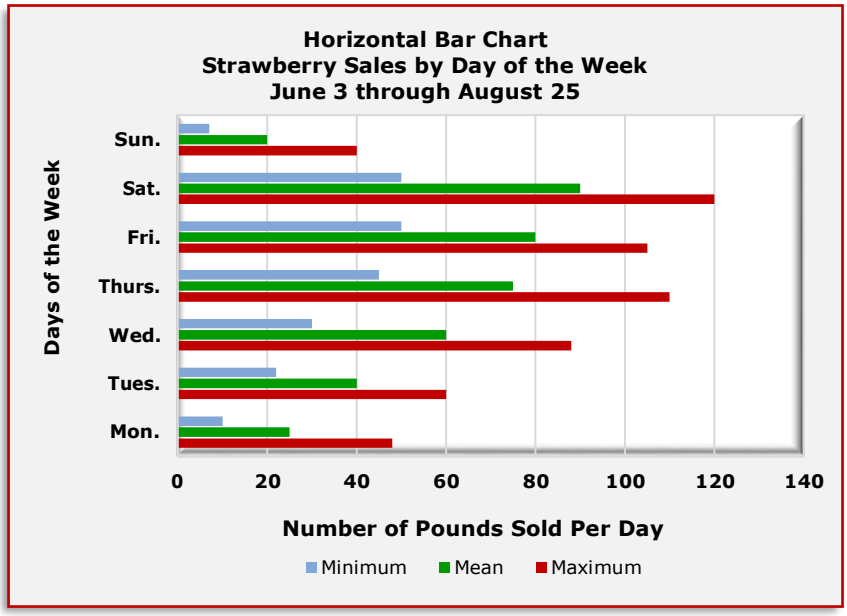
A bar chart is useful for comparing a statistic across groups. The height of the bar or the length of it, if the bar is displayed horizontally, displays the value of the statistic.

The following bar chart shows the mean number of pounds of strawberries sold per day over the twelve-week period. Thus, the Monday sales figure is the average of twelve Mondays, and so forth for each of the days of the week. This chart can be used to easily visualize which are the heaviest days of the week for selling strawberries in order to place orders at the appropriate times.

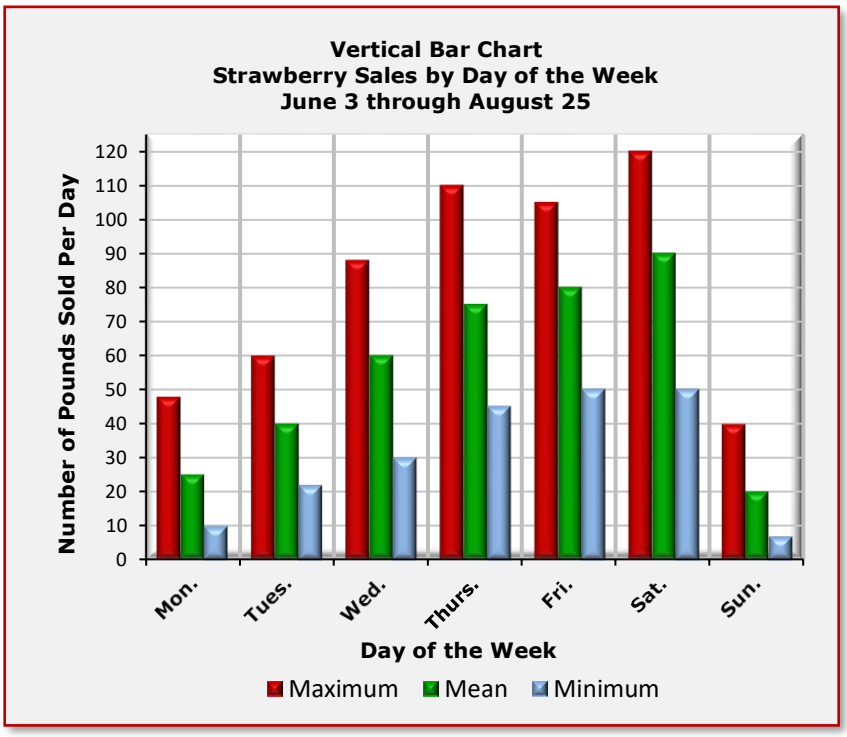


A bar chart can also be used to portray values horizontally, and when it does, it becomes the exception to the rule that the independent variable is generally on the *x*-axis (horizontal) and the dependent variable is generally on the *y*-axis (vertical). When a bar chart portrays the values horizontally, the independent variable is on the vertical axis and the bars extend to their values on the horizontal axis, representing the dependent variable.

The following horizontal bar chart is used to show not only the mean sales in pounds for each day of the week but also the minimum sales and maximum sales for each day, which are also important information.



Here is the same information presented vertically:

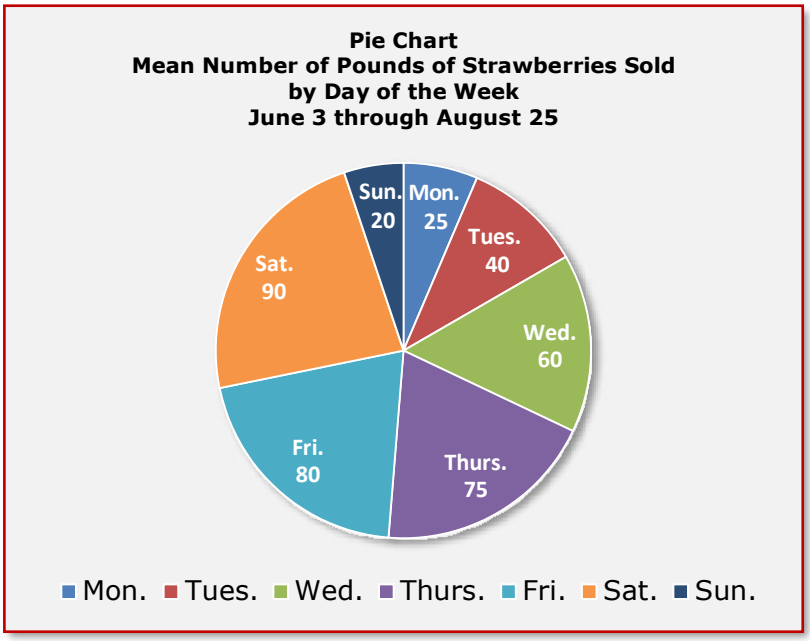


Pie Chart

A pie chart is in the form of a circle that portrays one value for each category, marked as pieces of a pie. In the example that follows of the mean pounds of strawberries sold per day for each day of the week over the twelve-week period, the pieces of the pie represent the days of the week and each one is sized to represent the mean sales for that day. The size of the "pieces" helps the user to visualize the relative sizes of the mean sales for each day.

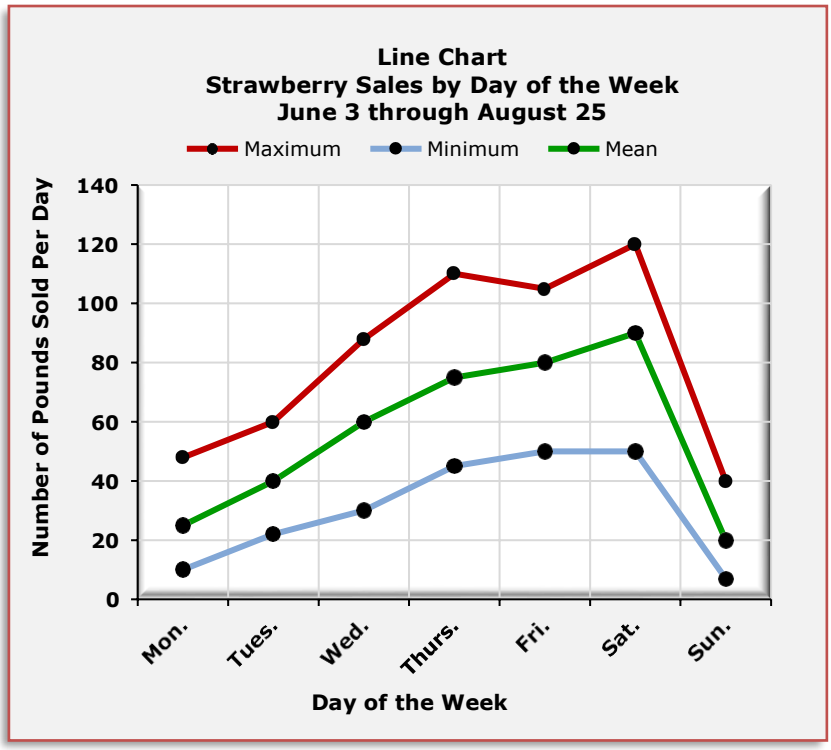
A pie chart does not have axes. Therefore, including values on the chart as labels helps users to interpret and use the information. The mean pounds of strawberries sold on each day have been added to each pie piece in the chart that follows. Values can be added as labels to data in other types of charts as well, but in some charts, doing so tends to make the chart hard to read.

A limitation of the pie chart is that it can present only one value for each category.



Line Chart

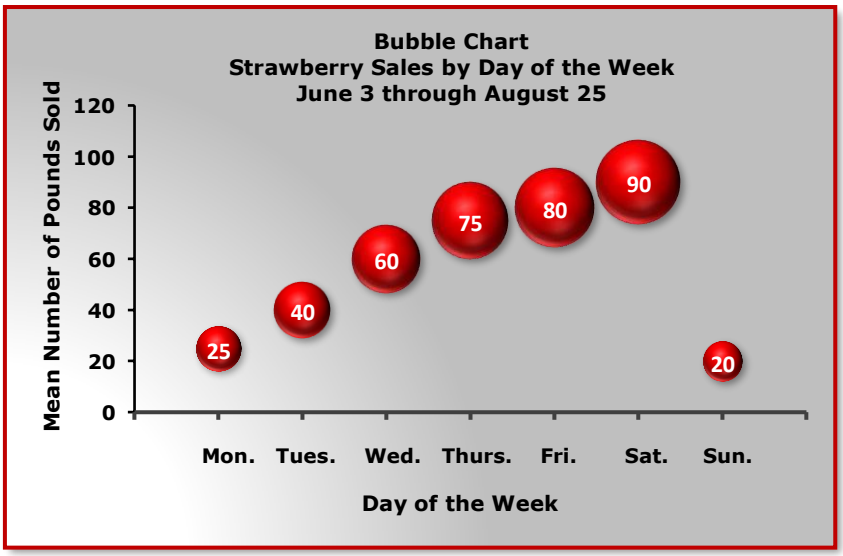
A line chart can be used to visualize several observations for each category, using one line for each observation. The strawberry sales data are shown on the following line chart as the minimum, the maximum, and the mean values for each day of the week for the twelve-week period.



Bubble Chart

A bubble chart replaces data points with bubbles that vary in size according to the size of the values they depict, thus adding an additional dimension to the chart: the relative sizes of the values plotted on the chart.

The following bubble chart shows the means of the strawberry sales for each day of the week.



Charts Containing the Entire Distribution of Values

Although summaries and averages are very useful, much can be gained by looking at additional statistics such as the median of a set of data or by examining the full distribution of the data. Histograms and boxplots can be used to display the entire distribution of a numerical variable.

Histogram

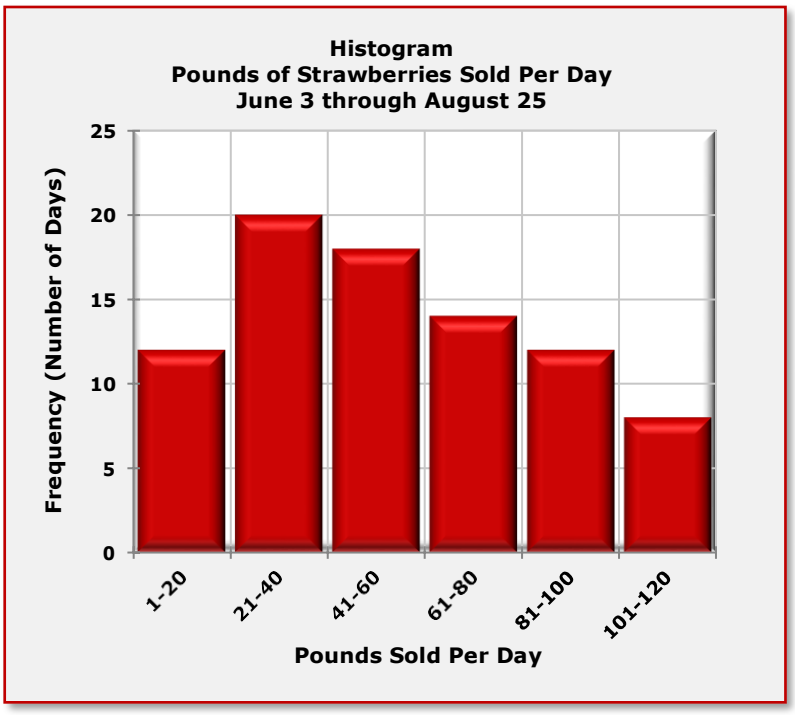
A **histogram** shows the frequencies of a variable using a series of vertical bars. The values of the variable may occur over a period of time, or they may be as of a moment in time.

The following histogram shows how many days during the twelve-week period the sales of strawberries were between 1 and 20 pounds, how many days between 21 and 40 pounds were sold, and so forth.

A histogram looks similar to a bar graph. However, it is different from a bar graph in that a bar graph relates two variables to one another, whereas a histogram communicates only one variable.

To construct a histogram, the range of values of the variable must be first divided into intervals, or **bins**, and then the number of values that fall into each interval are counted. For the histogram that follows, six bins are used. The bins and the values in each bin are:

Number of Pounds Sold Per Day	Frequency (Number of Days)
1-20 pounds	12
21-40 pounds	20
41-60 pounds	18
61-80 pounds	14
81-100 pounds	12
101-120 pounds	8



Boxplot

A boxplot is another type of chart that is used to display the full distribution of a variable. A boxplot shows the **minimum**, the **first quartile**, the **median**, the **mean**, the **third quartile**, and the **maximum** for each day of the week, as well as individual observations for each day of the week.

Note:

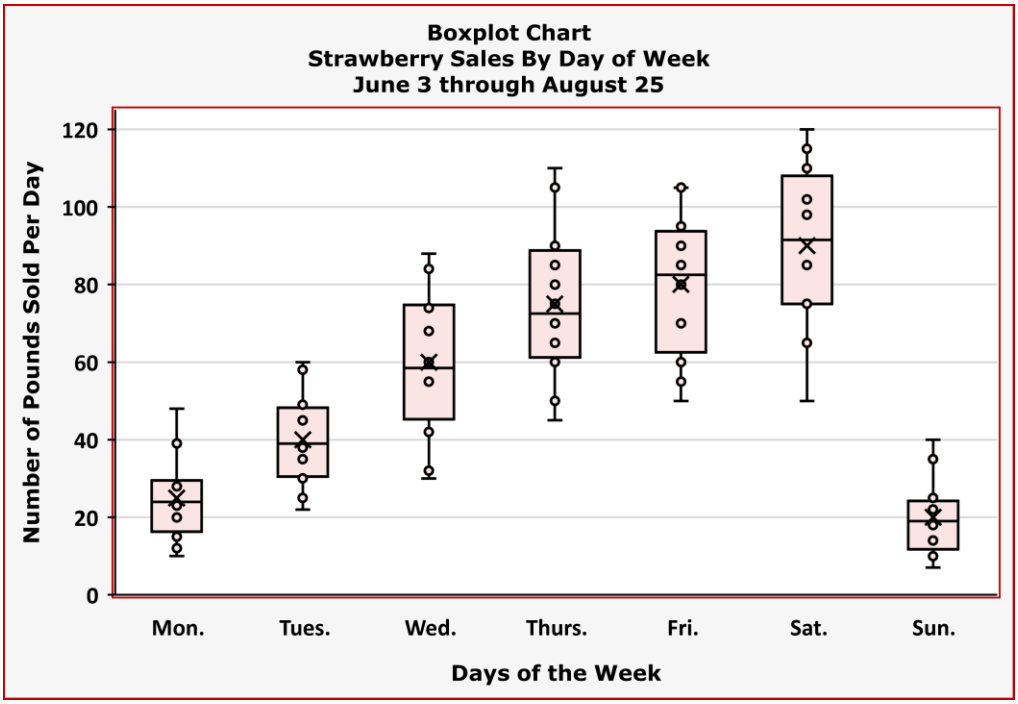
- The **minimum** is the smallest value in a distribution.
- The **median** is the middle value in a distribution. If the distribution contains an odd number of values, the median is the value with an equal number of values below it and above it. If the distribution contains an even number of values, the median is the mean (average) of the middle two numbers.
- The **first quartile** (**Q_1**) is the middle value between the minimum value and the median of a distribution; or, if two values are in the middle, it is the mean of those two values. If the median falls between two values, the first quartile is calculated as one-quarter of the difference between the middle number (or the mean of the two middle numbers) and the next **largest** value when the values are ordered from the smallest to the largest.
- The **second quartile** (**Q_2**) is the same as the median of the distribution.
- The **third quartile** (**Q_3**) is the middle value between the median and the maximum value in the distribution; or, if two values are in the middle, it is the mean of those two values. If the median falls between two values, the third quartile is calculated as one-quarter of the difference between the middle number (or the mean of the two middle numbers) and the next **smallest** value when the values are ordered from the smallest to the largest.
- The **maximum** is the largest value in the distribution.

In the table that follows, the data on daily strawberry sales for the twelve weeks have been re-ordered by day from the smallest value to the largest value for that day. Ordering the data in that way makes apparent the minimum, median, and maximum values for each day of the week and the approximate values for the first and third quartiles.

Note: The first quartile and the third quartile are **approximately** equal to the values in the table columns marked $\approx Q_1$ and $\approx Q_3$ that follow. The first quartile for each day is the middle value between the minimum value for that day and the median. Since the median is halfway between two values in this dataset, one-quarter of the difference between the number in the column marked $\approx Q_1$ and the next largest value is **added to** the number in the column marked $\approx Q_1$ to calculate the actual first quartile. The third quartile is the middle value between the median and the maximum value. One-quarter of the difference between the value in the column marked $\approx Q_3$ and the next smallest value is **subtracted from** the number in the column marked $\approx Q_3$ to calculate the actual third quartile.

Day of the Week	Mini-mum		≈Q_1			Median=the mean of these two columns				≈Q_3		Maxi-mum	Mean (calcu-lated)	Median (calcu-lated)
Mon.	10	12	15	20	22	23	25	28	28	30	39	48	25	24.0
Tues.	22	25	30	32	35	38	40	45	46	49	58	60	40	39.0
Wed.	30	32	42	55	55	57	60	68	74	75	84	88	60	58.5
Thur.	45	50	60	65	65	70	75	80	85	90	105	110	75	72.5
Fri.	50	55	60	70	80	80	85	85	90	95	105	105	80	82.5
Sat.	50	65	75	75	85	85	98	100	102	110	115	120	90	91.5
Sun.	7	10	11	14	18	18	20	20	22	25	35	40	20	19.0

A boxplot chart of the data follows.



Each day's box encloses the first quartile through the third quartile of values for that day. For example, on Wednesday, the bottom of the box is at the calculated first quartile of 45.25, calculated as 42 + (0.25 × [55 − 42]). The top of the box is at the calculated third quartile of 74.75, calculated as 75 − (0.25 × [75 − 74]).

The horizontal bar through each box marks the location of the median, which for Wednesday is the average of 57 and 60, or 58.5.

The "X" in each box marks the location of the mean for that day, which for Wednesday is 60.

The lines that extend vertically above and below each box with horizontal lines at the ends (called "whiskers") mark the minimum and maximum for each day. For Wednesdays, the minimum is 30 and the maximum is 88.

The circles mark values, and those outside the boxes are considered "outliers." "Outliers" are values that are far away from most of the other values in the distribution.